National Library of Canada
Acquisitions and Bibliographic Services Branch
395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale du Canada
Direction des acquisitions et des services bibliographiques
395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

Canada

University of Alberta

Reliability and Validity: A Model for Psychometric Analysis of Educational Tests in Zimbabwe

by

Esau Shingirai Nhandara

A thesis

submitted to the Faculty of Graduate Studies and Research

in partial fulfillment of the requirements for the degree of

Master of Education

Department of Educational Psychology

Edmonton, Alberta

Fall 1994

National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées.

The author retains ownership of the copyright in his/her thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without his/her permission.

L'auteur conserve la propriété du droit d'auteur qui protège sa thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

ISBN 0-315-94962-7

Canada

University of Alberta

Release Form

Name of Author:	Esau Shingirai Nhandara
Title of Thesis:	**Reliability and Validity: A Model for Psychometric Analysis of Educational Tests in Zimbabwe**
Degree:	M.Ed. in Educational Measurement
Year This Degree Granted:	1994

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly, or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as hereinbefore provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.

Nhandara
.....
(Student's signature)

71 Ruwa Avenue
Waterfalls
Harare, Zimbabwe

Date 25 august 1994

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled "Reliability and Validity: A Model for Psychometric Analysis of Educational Tests in Zimbabwe" submitted by Esau S. Nhandara in partial fulfillment of the requirements for the degree of Master of Education in Educational Measurement, Department of Educational Psychology.

Professor W. Todd Rogers, Supervisor

Professor E. W. Romaniuk

Professor Robert H. Short

Professor S. E. Sigurdson

Date: August 24/94

Abstract

The primary objective in this study was to develop a psychometric model for analysing educational tests in Zimbabwe. In order to do this, a 50-item multiple-choice geography test was pilot-tested using 495 boys and 414 girls in their second year of secondary education at 16 high schools in the Harare Region. The results, analysed in terms of the two main psychometric properties of testing—reliability and validity—have direct implications for testing programmes in Zimbabwe. The items were also reviewed by a 15-member validation panel for relevancy and representativeness.

The panel of judges agreed with the content classification with only four exceptions. However, as in other countries, there was less agreement on the cognitive skills being assessed by the test. The internal consistency of the full test was .79, which was considered close enough to the normally acceptable .80 for most norm-referenced tests. The subtests, as expected, had lower consistencies varying from .39 to .61. This was also in line with the generally expected increase in internal consistencies with increasing test lengths. However, there was one exception, mapwork, where the lower p-values resulted in a lower internal consistency measure.

With regard to discrimination indices, these were generally positive and above .20, with two exceptions, where they were negative. When combined with p-values, which generally averaged between .25 and .65, it was possible to identify test-wise-susceptible items. Test-wiseness is an artifact of item difficulty, and item writers have to develop a full complement of unflawed destructors.

In terms of the influence of gender, it was revealed that other factors intervene to compound gender differences. Such factors include the expectations of society, types of assessment instruments used, and syllabus emphases.

Each one of the factors cited above works to confound the valid interpretation of the results; that is, the test scores are open to misinterpretation. Accordingly, a series of recommendations for practice together with areas of additional study are identified.

Acknowledgements

I would like to extend my profound gratitude to my thesis supervisor, Dr. Todd Rogers, for his sterling support, encouragement, and guidance throughout this thesis. I also thank members of my thesis committee, Dr. Robert Short and Dr. Gene Romaniuk, for their comments and patience in editing this thesis; and Dr. S. E. Sigurdson for agreeing to serve as the external examiner.

As well, acknowledgements are offered to my wife, Bessie Fadzai, for assisting with data collection in Zimbabwe; and to Rose Moyo and Patricia Nhandara for providing much-needed stationery.

The research study could not have been possible without the financial and material support from the Ministry of Education and Culture, Harare, Zimbabwe; and the Zimbabwe-Canada General Training Facilities in both Ottawa, Canada, and Harare, Zimbabwe.

My special thanks are also extended to the 909 boys and girls from the 16 Harare Region schools who had to endure 1-3/4 hours of writing the examination, their teachers, headmasters, and Regional Office staff for their support and co-operation during the collection of the research data.

Table of Contents

Chapter — Page

1. Introduction: The Problem 1
   - Statement of the Problem and Rationale for the Study 1
   - Objectives Investigated and Questions Answered 2
   - Overview of the Context of Testing 3
     - Brief Historical Account of the Testing Scene in Zimbabwe's Schools 4
     - Educational Changes Following Independence in 1980 6
     - Localisation of Ordinary-Level Examinations and Its Impact 7
   - Potential Significance of the Study 11
2. Review of Related Literature 13
   - General Introduction 13
   - Distinction Between Reliability and Validity 14
   - Issues Pertaining to Reliability 15
     - Meaning of Reliability: A Conceptual Framework 15
     - Measurement Error and Its Estimation 16
     - Different Coefficients of Reliability 19
       - Equivalent-Forms (Alternate/Parallel Forms) Reliability 20
       - Test-Retest Reliability (Also Known as Coefficient of Stability) 21
       - Internal Consistency Measures of Reliability (Split-Half Method) 21
       - Internal Consistency of Reliability (K-Split Method) . . . 22
     - Interpretation of Reliability Coefficients 24
     - Unidimensionality 28

Chapter Page

Critical Review of Reliability in Educational Measurement 30

Issues Pertaining to Validity 31

Role and Meaning of Validity in Educational Measurement 32

Construct-Related Evidence 33

Content-Related Evidence 34

Criterion-Related Evidence 35

Validity Generalisation 35

Differential Prediction 36

Validity and Test-Wiseness 36

Gender Issues in Performance 38

Current Thinking on Validity 44

3. Research Methodology and Procedures 51

Instruments 51

The Sample of Examinees 53

Administration of the Test 53

Scoring and Data Preparation 55

Procedures for the Analysis of Results 55

Internal Consistency Indices 55

Item Analysis Using Classical Test Score Model 56

Point-Biserial and Biserial Correlations 56

Point-Biserial Correlation 56

Biserial Correlation 60

Estimating Item Discriminating Power (D) 61

Item Analysis Using Item Response Theory 62

Chapter Page

Item Parameters in Item Response Theory (IRT) 62

Issues of Validity 67

Content and Cognitive Representativeness 67

Test-Wiseness 68

Gender Differences 69

4. Presentation and Analysis of Results 70

Description of Students' Performance 70

Description of Test Characteristics 75

Internal Consistency 75

Relationship Among Subtests 76

Item Characteristics 77

Item Difficulty 77

Item Discrimination 77

Application of IRT to the ZJC Geography Test 80

Test Characteristic Curve 81

Standard Errors of Measurement 81

Item Fit Statistics 82

Description of Sources of Validity Evidence 83

Item-Topic Classification 83

Content Representativeness 86

Cognitive Complexity 86

The Influence of Test-Wiseness 87

Evidence of Gender Differences 90

Chapter Page

5. Discussion and Conclusions 95

Interpretation of Findings 95

Reliability and Item/Test Difficulty 95

Item Discrimination and Distractor Effectiveness 97

Content Representativeness 98

The Presence of Test-Wiseness 101

Gender Differences in Performance 102

Test-Score Use and Interpretation 104

Implications for Practice in Zimbabwe 107

Limitations of the Research 108

Recommendations for Further Study 109

References 111

Appendixes:

A. Glossary of Educational Measurement Terms 117

B. Zimbabwe Junior Certificate Geography Specimen Paper 124

C. Instruction Sheet to Judges, Zimbabwe Junior Certificate Geography Specimen Paper Specifications Table 142

D. Major Psychometric Properties of the Test Items 145

E. An Outline of Test-Wiseness Principles 153

List of Tables

Table		Page
1.	Summary of Approaches to Reliability Estimation	27
2.	All GCSE Groups' Entries 1988-90	41
3.	All GCSE Groups' Percentage Grades A-C, 1988-90	42
4.	GCSE Mathematics Entry Patterns 1988-90	43
5.	Comparison of Gender Differences (Female-Male) for 1985 GCE and for 1988 GCSE	45
6.	Types of Schools in the Harare Region (1993) Included in the Test Sample	54
7.	Summary of Student Performance on Four Subtests and Total Test	71
8.	Correlations Between Subtest, Total Test, and Gender	76
9.	Item Difficulty Levels	78
10.	Item Fit Statistics for Items 5, 34, and 41	83
11.	Item Content Classification by Judges	85
12.	Items Identified as Being Susceptible to Test-Wiseness	88
13.	Distribution of Test-Wise Items According to Test-Wise Cue Used, Items Showing Chance, and Bad Items	88
14.	Distribution of Test-Wise Items by Subtest	89
15.	t-tests for Independent Samples of Sex	91
16.	Summary of Differences Between Boys' and Girls' Performance	92
17.	Gender Differences at the Subtest Level	94

List of Figures

Figure		Page
1.	Expected Reliability *(r)* of a Test of Size *c* as a Function of the Number of Alternatives *a*	25
2.	Facets of Test Validity	48
3.	Characteristics of a Highly Discriminating Test Item	58
4.	Characteristics of a Moderately Discriminating Test Item	58
5.	Characteristics of a Nondiscriminating Test Item	59
6.	Characteristics of a Negatively Discriminating Test Item	59
7.	Trace Line for a Single Item	65
8.	ICC of Two Similar Items of Different Difficulty	65
9.	ICCS for Two Different Items	66
10.	Mapwork Scores	72
11.	Physical Geography Scores	72
12.	Economic Geography Scores	73
13.	Population and Settlement Scores	73
14.	Total Test Scores	74
15.	Plot for the Easiest Item	79
16.	Plot for a Moderately Difficult Item	79
17.	Plot for the Most Difficult Item	79
18.	Test Characteristic Curve for the 50 Items on the ZJC Geography Test	82
19.	ICC for a Very Difficult Item Which Also Discriminates Poorly (Item 5)	84
20.	ICC for a Moderately Difficult Item Which Discriminates Poorly at Higher Ability (Item 34)	84
21.	ICC for an Easy Item Which Discriminates Poorly at the Highest Ability (Item 41)	84

Chapter 1

Introduction: The Problem

Statement of the Problem and Rationale for the Study

The primary objective of this study was to provide a psychometric model for both the development and the analysis of educational tests in Zimbabwe. It is pertinent to observe here that, although the development of national or public school tests in the country has incorporated what can be claimed to be sound educational and achievement testing practices such as the use of specification tables and assessment objectives, none of these claims has really been confirmed in any research study. As a result of the combined effect of the lack of requisite funding for research and the absence of relevant research expertise, it has not been possible to carry out any research exercise in the assessment of students who write public school examinations. In addition, the proper research climate did not exist until recently, as will be illustrated in the section on the background to the problem presented below. In the context of Zimbabwe, therefore, examinations or public achievement tests continue to be administered to pupils at various points of schooling (end of primary school, end of first two years at secondary, end of fourth-year secondary, and end of sixth-year secondary) under the express understanding that the public examinations are all technically accurate and professionally sound despite the persistent absence of any research evidence that would have confirmed this belief or otherwise. Thus the most important single record of a pupil's performance has never been subjected to any regular and meaningful test to see whether what has been claimed by the test developers is in fact the truth. Fundamental questions such as the cognitive levels and processes which test constructors assume are used by students when answering questions have never been questioned. The concept of unidimensionality as reflected in the reliability coefficients and test scores has never been probed, nor has validity of score interpretation. In summary, the basic question of whether the decisions made

about thousands of school-leaving children in Zimbabwe on the basis of their performance in national tests are necessarily accurate and valid has not been addressed. There was therefore an immediate need to address the two fundamental but closely related questions of validity and reliability in educational testing in Zimbabwe.

Objectives Investigated and Questions Answered

In carrying out the study, there were certain specific issues that had to be investigated and questions that had to be answered in order to provide a meaningful psychometric model. To do this, the 1987 ZJC Geography Paper 1 specimen paper was used. Briefly, the test which was administered to students in the Harare Region during the period June-August 1993 contained 50 five-option multiple-choice items organised in terms of four subtests corresponding to the four main topics in the ZJC syllabus.

Under the issue of reliability the researcher aimed to establish the extent to which test items work together in a total test. Of particular importance was the concept of the internal consistency of the test. Were the different test items measuring the same content domain? Was there some kind of continuum in the ability and process that were being tested by the individual subtests in the test assessed, or were there any noticeable differences among the reliability coefficients of the different subtests? Did reliability coefficients increase with an increase in the number of test items as was claimed by relevant literature on test reliabilities?

Validity is regarded as the main psychometric consideration in educational testing, and crucial aspects which had to be probed under it included the balance between subject domain and skill representativeness through the analysis of specification tables, differences in the level of difficulty of test items, the effects of

items susceptible to test-wiseness on validity, and possible differences in performance resulting from gender differences.

Overview of the Context of Testing

In virtually every country there is created a sense of national crisis in education. The mood is one of the need to reform an outdated, irrelevant, or even mediocre education system. In the United States, for example, the 1983 National Commission on Excellence, in its report *A Nation at Risk*, warned the nation of its mediocre education system (Harris & Pickle, 1992). The 1988 Education Reform Act in the United Kingdom ushered in a National Curriculum and Unified Assessment System through the General Certificate of Secondary Education, popularly known as the GCSE (Stobart, Elwood, & Quinlan, 1992). In Zimbabwe, the somewhat frenzied interest shown in education following that country's independence in 1980 resulted in the New Content and Structure of Education (Government of Zimbabwe, 1987; Masango & Nembaware, 1991; Nhandara, 1993). The main thrust of this education-reform movement has been the need to introduce unified national curricula and assessment instruments (Ministry of Education and Culture, 1987a, 1987b).

Cole (1984) and Shepard (1991) have suggested that much of the educational innovation that has been taking place from the mid 1980s has involved the field of testing. Consonant with this change, a whole plethora of assessment vocabulary has been created, examples of which include "authentic assessment," "direct assessment," "performance assessment," "basic skills testing," "minimum competency test," and even "customised testing." Such a development has not been surprising, given the critical role that has been accorded education and testing throughout the development of countries. Education has indeed been viewed as the panacea of all our economic and social ills. Policy makers believed that education, through public examinations, enabled them to make the correct selection decisions, to award scholarships amidst

limited resources, to reward everybody through certification, to improve the quality of teaching and learning, and to foster a healthy sense of competition between institutions and individuals (Cole, 1984).

Testing has been one major way through which authorities have assessed educational accountability. Financial resources have often been awarded on the basis of examination results at institutions. Such decisions, where performance in examinations is used as the sole criterion for funding schools, demonstrate the extent to which test scores have been misinterpreted.

Testing has sharpened the ways in which the basics in education have been defined. Given any subject area and its assessment objectives, varying degrees of subject domain specificities have been possible. This exercise has helped to clarify a number of problems in the testing field, especially the development of appropriate assessment instruments. Stress has been placed on the use of more than one type of assessment instrument and on the use of assessments spread over a period of time to produce more valid results than single-shot examinations coming at the end of specific curricula. One form of assessment currently receiving a lot of attention is performance assessment, where the examinees actually construct their responses in contrast to selecting a response from the options provided (as in a multiple-choice test item).

Brief Historical Account of the Testing Scene in Zimbabwe's Schools

Systematic professional testing started in earnest in 1986 with the establishment of the Test Development and Research Unit located at the Examinations Branch of the Ministry of Education and Culture. Before this date the only testing exercise that bore some semblance of professionalism existed at Grade 7 (end of primary school), where regional panels generated items which were then built into examination papers under the auspices of the Schools Psychological Services. Unfortunately, this system

of examining applied only to African children (whites, Asians, and coloureds did not write an official examination at the end of their primary school).

At the Zimbabwe Junior Certificate level (ZJC, which is written by students at the end of their second year of secondary school), the Examinations Branch commissioned individuals considered to possess expertise in their subject areas to set and mark examinations in a subject area. Such examiners often did not coordinate with each other when setting examination items, and they did not receive any training in setting and marking examinations. In contrast to the Grade 7 Examination, the ZJC Examination was run by individuals without any professional guidance from the Examinations Branch (there was nobody appointed to do this work until 1986).

At the Ordinary and Advanced Levels (end of fourth and sixth year of secondary education), the examinations were set and marked in the United Kingdom, except for Shona and Ndebele (the two local languages). Very often teachers had to imagine what the tests would look like, because there was no guidance in the official syllabuses. These teachers were provided with little more than a list of topics to be covered.

When Zimbabwe became independent in 1980, it found itself with nine different examinations catering for different races and with no meaningful expertise in the field of professional test development. A salient point to note here is that these assessments were designed so as to allow only a very small proportion of African children to proceed beyond Grade 7. For example, at independence only 12% of African children were allowed to proceed from the seven-year primary to the two-year secondary education. By contrast, non-African children could proceed from primary to secondary with minimum difficulty (they did not have to write a public examination until ordinary level).

Educational Changes Following Independence in 1980

The first task for the new government following independence in 1980 was to democratise access to education, especially at the secondary-school level. The rapid expansion in secondary-school enrolment is illustrated by a dramatic rise from a preindependence figure of 66,215 in 1979 to 661,361 in 1990 (Masango & Nembaware, 1991). The net effect of this expansion was that secondary education became available for everyone who could afford to pay for it.

Parallel with the rapid increase in enrolment was a major revision of the educational curriculum for various levels of school examinations. At the ZJC level the major task was to make the syllabuses more relevant to the needs of a newly independent country. At this time, the first real attempt was made to clarify to some degree the teaching and assessment objectives for the various ZJC syllabuses and to include specimen (sample) examination questions with the official syllabuses. Admittedly, a reasonable effort was made to refine the various educational syllabuses so that they conformed with the latest worldwide trends such as were taking place in the United Kingdom and North America. However, unfortunately for the ZJC this effort was hampered because greater effort was devoted to the ordinary-level syllabus revisions. In addition, the embryonic Test Development and Research Unit staff, unable to cope with the demands of both the ZJC and the ordinary level, gave greater attention to the ordinary level examination. The feeling existed in Zimbabwe that, although the ZJC was an important examination, it was more of an achievement test rather than a basic school-leaving examination, which the ordinary-level examination was.

Localisation of Ordinary-Level Examinations and Its Impact

Localisation of the ordinary-level examination in Zimbabwe, which started in 1984, was primarily aimed at establishing a self-sufficient examination system which would eventually transfer all responsibilities for administration, computerisation, and professional activities such as test construction, marking, and training of examiners from the University of Cambridge Local Examinations Syndicate (UCLES) in the United Kingdom to the Examinations Branch, Ministry of Education and Culture, in Zimbabwe.

The agreement between the government of Zimbabwe, UCLES, and funding organisations recognised that the localisation of these three testing activities could not occur simultaneously. Thus when the localisation exercise started in 1984, it began with the training of examination markers. The training of question setters followed in 1988. Computerisation and related administrative aspects were the last aspects to be tackled in 1990. For the most recent examination in November 1993, UCLES continued to provide computerised back-up, but the professional aspects of examining were handled almost entirely by Zimbabwean examiners with the assistance of an UCLES consultant (Howarth, 1991; Nhandara, 1993).

The localisation project in Zimbabwe also provided for a period of three months training at UCLES for officers from the Examinations Branch. Over the years a core of some 15 officers has undergone training at UCLES, with special emphasis on the professional aspects of examination processes such as item development, grading, grade reviews, and related computer applications in the development of a sound examination system.

The localisation project was viewed as an exercise that would lead to the following developments:

- the development of more relevant curricula through the indigenisation of syllabuses, thus making them more relevant to the world of work;

- the creation of a competent pool of educators through training as examiners and participation in the examining process;
- the development of unified educational standards resulting from the creation of common examinations at different levels of the school system;
- increasing the employability of ordinary-level graduates because their curriculum would have been made more relevant to the needs of Zimbabwe;
- creating a school-leaving graduate who was self-reliant because the new curricula emphasises the cultivation of self-reliant skills; and
- ensuring that the ZJC examination provided a meaningful foundation to ordinary-level syllabuses.

Localisation in Zimbabwe has provided experience with how the development, monitoring, and control of an effective assessment system have had to adapt to ever-changing circumstances. In view of these developments, fundamental questions of validity have had to be raised. When the view is taken that the development of the ordinary-level examination was initially targeted at the top 25-30% of the school-going population in the United Kingdom, and that in the context of Zimbabwe the ZJC examination was ordinarily expected to be taken by an even smaller proportion of some 12-20% of the African population, the issue of validity in testing could surely never be discussed without reference to the crop of entry vis-à-vis the overall implied question of "For whom is the examination meant?" (Masango & Nembaware, 1991; Nhandara & Chagonda, 1990).

In order to appreciate the situation better, it is helpful to refer to the Education Act of 1987, Section 56, which stated: "The Secretary shall determine curricula and examination systems for all schools and in so doing shall not determine different curricula and examination systems for different schools on the grounds that they are government or nongovernment schools." The significance of this section was that there should be one curriculum for all schools at the same educational level. The Act

also classified schools into government and nongovernment schools. Among the nongovernment schools were private or trust schools set up by companies, rich parents, or companies (syndicates) such as education board (local authorities) or church-related schools (Catholic, Methodist, Anglican, Presbyterian, Lutheran, and Salvation Army).

In order to devise and implement the requirements of Section 56, the Education Development Unit, made up essentially of the Curriculum Development Unit (CDU) and Examinations Branch (EB), was set up to take charge of school syllabuses and examinations, respectively. Both the CDU and the EB were staffed largely by former primary and secondary school teachers, virtually all of whom who had never received any professional training in the design and development of educational curricula, and certainly none of whom had ever received any training in the basics of examining. No college or university in the country offered any training in the basics of examining, although some institutions taught syllabus design at varying levels of detail.

The net effect of all the developments that went before independence and those that followed soon after it was that the syllabuses at the various school levels remained largely academic. It was these syllabuses, though in significantly altered form with respect to content, that were to become the only educational diet for every child of school-going age. The examination system was also unified into one examination for each level of the school population.

The view prevails in Zimbabwe that a labour-intensive system of manual marking as opposed to machine-scored objective testing is the only desirable way to ensure meaningful national examination standards. As of November 1993 some five multiple-choice papers existed at the ordinary-level examination. At the ZJC level there are 13 multiple-choice examination papers out of a total of about 44 examination papers. (Geography is one of the 13.)

The background has been made deliberately long and might appear irrelevant to the main research thrust at first sight, but the researcher's considered view has been that, given the nature of the testing scene now in existence in Zimbabwe, issues of validity could be better understood and interpreted only against this kind of background. Of particular concern arising out of this background, and having a direct bearing on validity, were the following:

- Uniform examinations such as those at the ZJC level now involve children from different experiences: on the one hand, African children who have had experience at both the Grade 7 and ZJC examination level; and, on the other hand, whites, Asians, and coloureds who have only recently started writing official examinations at both the Grade 7 and the ZJC level.
- The use of specification tables in constructing tests is a recent development which is not fully comprehended by many people, including the examiners, especially when it involved the classification of levels of cognitive skills.
- Multiple-choice testing is a recent development, and its use as an assessment instrument is still limited, at least in comparison with its almost universal application to the North American testing scene.
- Some of the multiple-choice-items-related concepts were virtually unknown to many examiners, let alone to the school children. Of particular concern was test-wiseness, which is so closely related to multiple-choice tests. In addition, official test results do not include the test reliabilities largely because of the limited computer-software packages and relevant professional expertise.
- The ZJC multiple-choice tests were not pretested before being administered to the ZJC candidates.

Arising out of the situation described in this section is the need to realise that official and professional testing is still very much in its infancy in Zimbabwe. Any trend or conclusion that could be identified or discerned is only a tentative indicator

of the possible way that things could have moved or could move in the future. Extreme care and caution have to be taken when drawing inferences from the test scores and other related psychometric measures with regard to the educational tests currently being administered to school children in Zimbabwe.

Potential Significance of the Study

By carrying out this study, the researcher hoped to be able to provide a practical psychometric model that would serve as a structure for evaluating educational tests in Zimbabwe. As already noted, no such framework exists in the country for use by both test developers and test users. In view of the fact that the model addresses the two fundamental issues in educational measurement—validity and reliability—it was also hoped that the testing programme, whose initial foundation was firmly and soundly established with the creation of the Test Development and Research Unit of the Examinations Branch, would have its base extended to include another milestone on the road towards the creation and development of a professionally sound testing system.

Given the proposal that certain fundamental decisions should be made using test scores from national examinations, the study should be of considerable interest to both test developers and test users. In particular, if the proposed move to channel children into the three categories of academic, science, and vocational streams becomes a reality in the near future, there is even more need to study the crucial issues of validity and reliability to shed more light on the accuracy of the test scores obtained by children in their ZJC examinations.

The research study was also viewed as a useful examination programme-evaluation exercise. No educational programme has remained static; rather, programmes have changed in response to social, political, and economic conditions. State legislators have prescribed minimum educational requirements according to the

political ideologies to which they subscribed. Educational-programme evaluations, of which this study is an example, are useful exercises in that they point out weaknesses and strengths of the programme being examined. In the case of the present study, the programme of interest is the testing system.

It has become a widely accepted fact that no process of planning and implementation is complete without evaluation. Evaluations have been undertaken as quality-control measures for management and administrative purposes, for planning and policy development, and to meet both fiscal-accountability and social-acceptability requirements.

Further, this research study, coming as it did at this particular time, and although somewhat opportune, was in many ways long overdue. The issues that were being probed now should have been explored before so that today the question "How better do we understand the issues?" and not "How should we aim to explore and understand the problems?" would be the focus of study.

Chapter 2

Review of Related Literature

General Introduction

In the field of educational measurement and assessment there are a number of key terms whose application and understanding call for specific meanings. A selection of such terms, considered appropriate to the present study, is provided in Appendix A. Additionally, it is also necessary to indicate in this section on the review of related literature with which of the two major types of testing currently available this study is concerned. This is done towards the end of the section below.

One controversial area in measurement today centres upon the debate between Norm Referenced Tests (NRTs) and Criterion Referenced Tests (CRTs). Whereas norm referenced testing concerns itself with determining at what level an examinee falls on a specific trait, where a trait is conceptualised as a characteristic which different people possess in different amounts, concern in criterion referenced testing lies in determining the degree or extent of mastery of a defined domain of knowledge or skill. An item on an NRT may look exactly the same as one on a CRT, but the interpretation given to examinee scores is completely different. NRT scores make possible the comparison of the performance of each examinee to that of the norm group across certain specified content coverage. Conversely, CRT scores show whether individuals or groups have met or have not met the criteria of mastery of specific educational objectives.

For the purposes of this study, the term *test score* is used in the context of norm referenced scoring only. Consequently, the review of related literature is limited to norm referenced testing only.

Distinction Between Reliability and Validity

All measurement specialists have agreed that the concepts of reliability and validity are central to the theory and practice of educational and psychological measurement and evaluation (Crocker & Algina, 1986; Cronbach, 1990; Hopkins, Stanley, & Hopkins, 1990; Mehrens & Lehmann, 1991; Sax, 1989; Thorndike, 1977). They have further contended that, although validity and reliability are closely related, they refer to different things. According to Ary, Jacobs, and Razavieh (1990) a decision as to whether we are dealing with validity or reliability depends on whether the measurement errors are systematic or random. For example, the scores of non-English subjects could be systematically depressed because the subjects did not comprehend what they were expected to do, or the test might have been culture specific to children from an English-speaking background. Such systematic errors of measurement are a validity problem. Random errors, on the other hand, are the opposite of systematic errors and constitute a reliability problem. Random errors arise from a number of sources. Errors might be inherent in the instrument itself. The test might be too short, so that those subjects who happen to know only the few answers required for the particular short test obtain higher scores than they deserve, whereas those who do not know those specific answers but several other different but related answers obtain lower scores than they deserve. A classic example could be a test given to assess how well students know a selection of five African capital cities and their countries, in which those who know only those particular five cities and their countries get everything, and those who know 40 others and their countries get lower marks. Such a situation results from the fact that, in a shorter test, chance is more of a factor than it is in longer tests.

Ambiguous test items also encourage guessing, resulting in a few "lucky" ones who might respond according to the intentions of the examiner achieving higher scores than do the "unlucky" ones who answer in another but equally correct way.

Inconsistent scoring procedures are also cited as a major factor in test reliability. Precise scoring procedures enhance reliability, whereas vague scoring procedures depress it. Random errors might also result from the administration of the instrument when inexperienced persons depart from standardised procedures in either test administration or scoring.

Testing conditions such as light, heat, and ventilation also affect performance and, ultimately, both the reliability and the validity of the test. Extraneous conditions such as the occurrence of a storm, an earthquake, the collapse of a roof or building nearby, and sudden fire warnings are all factors that could affect performance and, ultimately, both the reliability and the validity of the particular test.

Pupil error resulting from fluctuation in motivation, interest, fatigue, physical condition, anxiety, and other mental and emotional factors affect test scores and, ultimately, the reliability.

There is a real need, therefore, to keep a clear distinction between the issues of reliability and validity as they pertain to measurement. Otherwise, it is quite easy for one to discuss validity as if it were the same as reliability. For convenience, we glibly speak of the reliability and validity of a test, but it is more technically correct to speak of the reliability and validity of the test scores rather than of the test itself.

Issues Pertaining to Reliability

Meaning of Reliability: A Conceptual Framework

Feldt and Brennan (1989) regarded the quantification of the consistency and inconsistency in examinee performance as constituting the essence of reliability studies. To have done its job well, a measure therefore must have yielded accurate results (Hopkins, Stanley, & Hopkins, 1990). In other words, a test has little value if the score it yields for Gwinyai today is quite different from the score it would yield for him under similar conditions tomorrow (Crocker & Algina, 1986; Hopkins et al.;

Mehrens & Lehmann, 1991). On the other hand, it is quite possible theoretically for a test to yield consistent scores from day to day without the scores having any validity (Ary et al., 1990; Hopkins et al., 1990). What this means is that a test could yield consistent scores whose meaning and interpretation are still doubted. As an example, a scale on honesty could yield consistent results on different occasions, and yet the inferences drawn from the scores are considered invalid or untruthful. If we extend the argument further, this means that reliability alone has very limited value in the interpretation of the meaning which the test scores are given. In this regard reliability has to be related to the more important concept of validity for it to assume any usefulness in the interpretation of and inferences drawn from educational-test scores.

Estimates of the reliability of test scores are required to assess the amount of measurement error expressed in the form of the standard error of measurement. Sometimes decisions are taken on the basis of individual scores, but in the majority of cases individual scores or their aggregations are considered with reference to other scores. Test scores are often used to determine certification for successful completion of a course of study or otherwise, pass or fail, mastery or nonmastery, and a host of other situations. It is absolutely essential that when such decisions are taken, they are taken with reference to the standard error of measurement, especially at the cut score.

Measurement Error and Its Estimation

It is pertinent to observe at the outset that, no matter how much we try, any score that we obtain from a measuring instrument we use will always contain some error of measurement. These errors derive from a number of sources, such as nonstandardised test administration or scoring, and examiners' and examinees' errors. The errors are all collectively referred to as error of measurement and can mathematically be denoted as (E). The reliability, and therefore the dependability, of the test score is determined by the size of the error of measurement (E). The smaller

the measurement error, the more dependable and reliable the test scores are, and the more valid the inferences made from the test scores are likely to be (Ary et al., 1990; Ebel & Frisbie, 1991; Sax, 1989; Smith & Glass, 1987). In one sense, therefore, reliability refers to the degree to which test scores are free from measurement errors. Basic to the proper evaluation of any test are the identification of major sources of measurement error; the size of the errors resulting from the sources; the indication of the degree of reliability expected between pairs of scores under particular circumstances; and the generalisability of results across items, test forms, item raters and item administrators, and other psychometric aspects (*Standards for Educational and Psychological Testing*, 1985).

Using classical test theory,

A person j has a true score of tau (τ);

But

all we can measure is his observed score Xji on test form i designed to measure τj. It is assumed that Ei(Xji) = τj.

However, Xji contains measurement error Eji. These three aspects are related:

$$Xji = \tau i + Eji$$

Consequently, Ei (Eji) = 0 (Crocker & Algina, 1986). It is assumed, reasonably though, that Eji will change randomly each time we measure person j, and that the distribution of the measurement error for person j on an infinite number of tests is a normal distribution. The measure of the spread of the distribution Eji around the mean error for each person (zero) is called the standard error of measurement for person j (SEMj). Clearly, we would prefer that this standard error of measurement be as small as possible.

One consequence of educational measurement practice is that we cannot test a person over and over again to compute SEMj. Instead, by testing more than one person at a time, we could get a good estimate of SEMj. Using the group standard

error of measurement, here designated σe, it can be shown that if the group standard error of measurement were estimated to be close to zero, the SEMj for each person would also be close to zero, which implies small error of measurement or none at all in an individual score, Xji. On the other hand, if the group standard error of measurement is large, then at least some of the SEMj for each person will be large, which in turn implies that some of the Ejis will be large.

Measurement specialists have also shown that because the error of measurement is random in the population of individuals, $\sigma x^2 = \sigma x^2 + \sigma^2 E$, where

σx^2 is the variance of the population of observed scores on test X,

$\sigma \tau^2$ is the variance of the population of true scores, and

$\sigma^2 E$ is the group variance error of measurement.

Spearman (1904) considered the ratio

$$\frac{\sigma \tau^2}{\sigma x^2} = \frac{\text{true score variance}}{\text{observed score variance}}$$

$$= \frac{\text{true score variance}}{\text{true score variance + error of measurement variance}}$$

$$= \frac{\text{signal}}{\text{signal + noise (error of measurement)}}.$$

RELIABILITY

$$\rho_{xx} = \frac{\sigma \tau^2}{\sigma x^2}$$

where ρ_{xx} is the reliability.

Combining the result with the observation that $\sigma x^2 = \sigma^2 \tau + \sigma x^2$, we get the formula for the group standard error of measurement:

$$\sigma e = \sigma_x \sqrt{1-\rho_{xx}}$$

To obtain an estimate of σe, it is first necessary to estimate ρ_{xx}. Given that this is possible, the formula for the standard error of measurement (SEM) for a sample of subjects is expressed as

$$S_{EM} = S_x \sqrt{1-r_{xx}},$$

where

S_{EM} = the standard error of measurement

S_x = the standard deviation of test scores

r_{xx} = the reliability coefficient.

If an achievement test has a reliability coefficient of .84 and a standard deviation of 4, then the S_{EM} equals $4 \sqrt{1-.84} = 4 \sqrt{.16} = 1.6$.

In the case cited above, the standard error of measurement that amounts to 1.6 has been suggested as a good indicator of the individual errors of measurement. Supposing that an individual in the group tested had an observed score of 75, we could build the 68% or 95% confidence levels as follows:

68% C.I. Lower limit = 75 − 1.64 = 73.4

Upper limit = 75 + 1.64 = 76.6

95% C.I. Lower limit = 75 − (1.96 x 1.6) = 71.86

Upper limit = 75 + (1.96 x 1.6) = 78.14

The interpretation of the standard error of measurement in relation to an individual's score of 75 means that there is a 68% probability that the 68% confidence band 73.4 to 76.6 spans the individual's true score, whereas the same band for the individual's true score is widened to 71.86 to 78.14 using 95% confidence level.

Different Coefficients of Reliability

There are four major procedures for estimating coefficients of reliability. These are parallel forms; test retest; internal consistency-split halves; and internal consistence–K-splits, where K is the number of items.

Equivalent-forms (alternate/parallel forms) reliability. The equivalent or parallel forms reliability requires the use of two instruments (forms), X and Y, supposedly equivalent in both content and skills being tested. Further, it is assumed that

$$\mu x = \mu y$$

and

$$\sigma x^2 = \sigma^2 y$$

When these conditions are met, an estimate of the parallel forms reliability is given by the Pearson-Product Moment Correlation:

$$r_{XY} = \frac{N\Sigma XiYi - \Sigma XiYi}{\sqrt{[N\Sigma Xi^2 - (N\Sigma Xi)^2][N\Sigma Yi^2 - (\Sigma Yi)^2]}},$$

where

r_{XY} = Pearson r (product moment correlation)

Σ_X = the sum of scores in the X distribution

Σ_Y = the sum of scores in the Y distribution

Σ_{XY} = the sum of the products of paired X- and Y-scores

Σ_X^2 = the sum of the squared scores in X-distribution

Σ_Y^2 = the sum of the squared scores in Y-distribution

N = the number of paired X- and Y-scores (subjects)

When the tests are administered on the same day, the resulting coefficients are known as the coefficients of equivalence; but when it becomes necessary to administer the two forms on different days, the resulting coefficient is known as the coefficient of equivalence and stability. The equivalent-forms procedure is generally considered to provide the best estimate of reliability in academic and psychological testing. The biggest problem with this procedure, however, is achieving equivalence or parallelism between the two forms used.

Test-retest reliability (also known as coefficient of stability). The test-retest coefficients indicate the consistency of subjects' scores over time and are not normally appropriate for achievement tests. The formula used to compute the equivalent forms reliability is used to compute the test-retest reliability, where X refers to time 1 and Y refers to time 2.

Internal consistency measures of reliability (split-half method). Internal consistency measures were designed to determine to what extent all the items in a test measure the same thing. They require a single administration of the same test and have become extremely popular because they avoid the need to construct two parallel forms of a test or to administer the same test on two different occasions.

The "common" split-half reliability, which divides the test into two "parallel" halves, starts with the correlation between the scores on two half forms. This correlation is then adjusted to obtain an estimate of reliability for the full test using the Spearman-Brown prophecy formula (Spearman & Brown, 1910).

The computation of the split-half method is given by the following formula:

$$r_{\text{FULL TEST}} = \frac{2r_{½½}}{1 + r_{½½}}$$

where $r_{½}$ refers to the reliability of one half of the test.

Rulon (1939), Guttman (1945), and Flanagan (cited in Kelley, 1942) were concerned that it would not be possible to divide a test into two parallel halves. They developed, independently of one another, formulae for use with tests that were tau equivalent. However, the formulae, which are presented below, are algebraically equivalent:

$$r_{\text{FULL TEST}} = 1 - \frac{Sd^2}{Sx^2} \quad \text{(Rulon's formula)}$$

$$= 2\left[1 - \frac{Sa^2 + Sb^2}{Sx^2}\right] \quad \text{(Guttman-Flanagan procedure)}$$

where

Sd^2 = the variance of the difference scores

Sx^2 = the variance of the total observed scores

Sa^2 = the variance of one half of the tests

Sb^2 = the variance of the other half test

Measurement specialists have discouraged the use of the split-half method with short tests, preferring instead to use it with relatively long tests. It is also not recommended for speeded tests, because it tends to yield spuriously high coefficients.

Internal consistency of reliability (K-split method). The K-split internal consistency measures are designed to estimate interitem consistency or homogeneity, thus ultimately acting as an indication of the extent to which the test items measure the same construct. Internal consistency measures reflect two major sources of measurement error: (a) content sampling or representativeness, and (b) heterogeneity of the domain sampled. The more heterogeneous the domain is, the lower the interitem consistency becomes; and, conversely, the more homogeneous the domain is, the higher the interitem consistency is.

Three coefficients of internal consistency are normally computed. The first, K-R20, can be used only with items that are dichotomously scored. There is no restriction on the form of scoring for the remaining two, Cronbach's alpha and Hoyt's ANOVA.

(a) K-R20: Reliability $r_{xx} = \frac{K}{K-1}\left(1 - \frac{\Sigma pj(1-pj)}{Sx^2}\right)$

where

K = number of items

Sx^2 = variance of scores on the total test (squared standard deviation)

pj = the proportion of examinees who answer item j correctly.

(b) Coefficient alpha (Cronbach, 1951):

$$(r_{xx}) = \alpha = \frac{K}{K-1}\left(1 - \frac{\Sigma Sj^2}{Sx^2}\right)$$

where

K = number of items

Sj^2 = the variance on item j

Sx^2 = the variance of the total test observed.

For dichotomously scored items, coefficient alpha and K-R20 (Kuder Richardson) are equivalent.

(c) Another estimate of internal consistency is Hoyt's Estimate of reliability (Hoyt, 1941), which uses a pxi (persons-by-items) random effects. The internal consistency is estimated by:

$$M_{xx} - 1 - \frac{MSpxi}{MSp}$$

where MSp is the mean square for person, and

MSpxi is the interaction terms between persons and items and items plus, because there is only one observation included, the residual error.

Hoyt's estimate of reliability yields the same value as Cronbach's alpha. In the case of a test battery composed of several subtests, the reliability of the aggregated or composite score is determined using Cronbach's composite for alpha or stratified alpha (Cronbach, 1965). Cronbach's composite for alpha

yields a lower estimate of reliability than does either Hoyt's estimate or K-R20 because its computation is based on the subtests and not on the individual items.

Today, it is no longer necessary to calculate the various reliability coefficients manually because there are several powerful computer programmes which can do so within a matter of seconds.

Interpretation of Reliability Coefficients

There are a number of basic considerations which have to be taken into account if the interpretation of coefficients is to become more relevant and meaningful; otherwise there exists a real danger that the studies on reliability will remain superficial. Chief among these considerations are the length of the test, number of alternatives in a multiple-choice item, group heterogeneity, individual abilities, the specific technique used to estimate reliability, and the nature of the variable that is being measured.

According to Ary et al. (1990), Beck (1974), Grier (1975), and Smith and Glass (1987), reliability is essentially a result of the combined effect of the five major factors outlined in the paragraph above.

1. The reliability of a test is in part a function of the length of the test. The longer the test is, the higher the reliability tends to be (see Figure 1). This is primarily because a longer test is more representative, because it samples the domain universe better. In fact, if it were possible to sample the entire domain universe, the individual scores would essentially be the true scores. One condition, though, should apply in the case above: all the items in the test should belong to the same domain universe.

2. As shown in Figure 1, the reliability is also influenced by the number of alternatives. Further, there is an interaction between test length and the number of

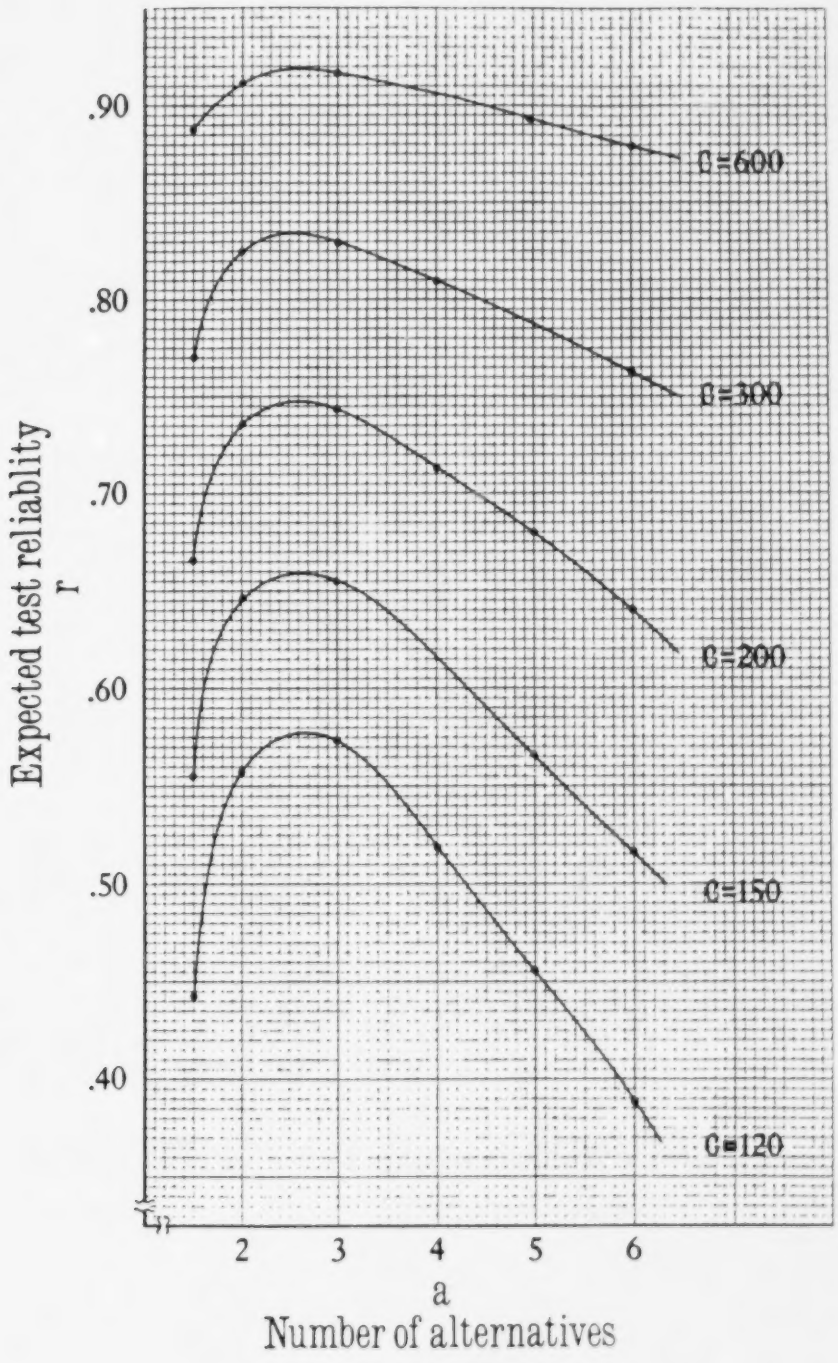

Figure 1. Expected reliability *(r)* of a test of size *c* as a function of the number of alternatives *a* (Grier, 1975, p. 111).

alternatives, with the difference in the value of the reliability coefficient decreasing with increasing test length.

3. Reliability is in part a function of group heterogeneity. The greater the variance in a sample is, the higher the reliability is. Conversely, samples made up of individuals who are all alike in the amount of the indicator they possess produce lower estimates of reliability.

4. The reliability of a test is in part a function of the ability of the individuals who take the test. The difficulty level of tests affects their reliability. When a test is difficult, subjects tend to guess a lot, resulting in lower reliability coefficients. When a test is too easy, all subjects score at the same level, resulting in low reliability coefficients because the items do not discriminate among the subjects. A test should be constructed to take into account the true range of differences among the subjects included in the test sample. "Reliabilities will be artificially low if the test or procedure does not have a high enough 'ceiling' (maximum score) or a low enough 'floor' (minimum score) to accommodate the true range of differences among the members of the sample" (Smith & Glass, 1987, p. 106). Feldt (1993) argued for a balance between easy and difficult items in a test:

> Examinees who are below average in ability may become discouraged when they encounter no exercises that they can answer with confidence in the first few minutes. As their discouragement mounts, concentration and commitment deteriorate. However, it may be possible to maintain the examinees' motivation if they meet with some success in the early portion of a test. (p. 48)

5. Reliability is in part a function of the specific technique used for its estimation. Different procedures for estimating the reliability of tests result in different coefficients of reliability. As shown in Table 1, and summarised by Ary et al. (1990), the equivalent-forms technique gives a lower estimation of reliability compared to either the test-retest or split-half procedures, because in the equivalent-forms technique form-to-form as well as time-to-time fluctuation are present. The

Table 1

Summary of Approaches to Reliability Estimation

Design	Coefficient	Major Sources of Error
Test-retest	Stability	Differences in people across time
Parallel/alternate forms		
- same time	Equivalence	Differences between forms
- different times	Equivalence and stability	Differences between forms and in people across time
Rulon/Guttman/Flanagan	Internal consistency	Differences between forms
Internal consistency "split into K items" (i.e., treat each item) e.g., K-R20 Cronbach's alpha	Internal consistency	Differences among items
- Split into two parallel halves		Differences between half forms
- Spearman-Brown correction		

(Crocker & Algina, 1986, p. 141)

split-half procedure, on the other hand, results in higher reliability coefficients than do its alternatives, because it is not sensitive to time-to-time fluctuations.

6. Reliability is in part a function of the nature of the variable that is being measured. Most established achievement and psychological tests yield consistently high reliabilities, whereas tests of personality traits yield only moderate reliabilities. Whether a reliability coefficient is considered high enough depends primarily on the use to which the test scores are to be put and the relationship between the obtained coefficients to existing competing tests. As Ary et al. (1990) observed, a spelling test with a reliability of .80 would be considered unsatisfactory if competing spelling tests have reliabilities of .90 or higher. An achievement test of geography with a coefficient of .80 would be considered excellent if existing competing tests have coefficients of only .60.

For research purposes coefficients as low as .30 to .50 have often been considered acceptable, but where important and irreversible decisions were to be made about individual people, only tests with reliabilities of .85 and over are considered acceptable.

Unidimensionality

The concept of *unidimensionality* refers to the situation in which an item or a test measures a single latent ability or trait. In practice, however, it is not possible to achieve total unidimensionality because several cognitive, personality, and test-taking factors affect test performance. According to Hambleton, Swaminathan, and Rogers (1991), what is required for the unidimensionality assumption to be met adequately by a set of test data is the presence of a dominant component or factor that influences test performance. This dominant component or factor is then referred to as the ability measured by the test. However, it was noted that the ability measured is not necessarily inherent or unchangeable. Ability scores are expected to change over time as a result of learning, forgetting, and other factors.

Another concept that is closely related to unidimensionality is *local independence*, an assumption that responses an examinee makes to different items on a test are statistically independent. In other words, an examinee's performance on one item must not be affected in any way by performance on any other item in the test—no items should provide clues to the answers of other items. Whenever the assumption of unidimensionality is true, local independence is obtained, and the two concepts become equivalent. Local independence could be obtained, however, even when the data was not unidimensional, such as when all the ability dimensions influencing performance have been taken into account. Conversely, local independence cannot be achieved where the complete latent space has not been specified, such as in a geography test that requires, in addition to the basic

geographical skills, a lot of reading. Examinees with poor reading skills will not answer the items correctly regardless of their geographical proficiency. A dimension other than geographical proficiency, that is, reading skill, will influence their performance. On the other hand, if all examinees have the requisite reading skills, only geographical proficiency will influence their performance, and local independence will obtain. Similarly, the ability by examinees to detect clues (test-wiseness) might benefit some examinees more than others, thereby violating the principle of local independence.

In practical situations it is most unlikely for several test items to be uncorrelated, which means that an examinee's responses are also correlated. What the concept of local independence means in such situations is that when the common traits are "partialled out" or "held constant," the variables being measured become uncorrelated, thereby satisfying the concept of local independence.

From a psychometric point of view, the concept of internal consistency in test data is reflected in the reliability coefficients—the higher these are, the more homogeneous the test is. The importance of homogeneity in educational testing has been aptly summarised by McNemar (1946; cited in Lumsden, 1976):

> Measurement implies that one characteristic at a time is being quantified. The scores on an attitude scale are most meaningful when it is known that only one continuum is involved. Only then can it be claimed that two individuals with the same score or rank can be quantitatively and, within limits, qualitatively similar in their attitude towards a given issue. As an example, suppose a test of liberalism consists of two general sorts of items, one concerned with economic and the other with religious issues. Two individuals could thus arrive at the same numerical score by quite different routes. Now it may be true that economic and religious liberalism are correlated, but unless highly correlated the meaning of scores based on such a composite is questionable. (p. 266)

Critical Review of Reliability in Educational Measurement

Lumsden (1976), in his critical review of the classical test theory, has concluded that he could only find three propositions about reliability worth remembering by test users and test constructors. The three propositions are:

1. Test scores are unreliable because performance fluctuates from time to time depending on a number of factors that have already been covered under reliability.

2. All other things remaining equal (item type, dispersion of item difficulties, item correlations), and up to a certain point, a longer test is better than a shorter test because it tends to sample the domain universe better and always results in higher reliability coefficients. But the fact was implied that at a certain point very often fatigue and boredom intervene to affect item functioning.

3. A test with a higher item consistency (intercorrelations)is better than a test with lower item intercorrelations.

Lumsden (1976) has urged that the study of reliability in its present form (classical test score model) should be abandoned in favour of the study of generalisability theory and validity because these two notions are more relevant and also accurate with regard to educational testing. It should be noted that the generalisability theory is somewhat more complex than classical test score reliability theory, and that generalisability theory subsumes classical reliability (Crocker & Algina, 1986).

Ary et al. (1990) and Feldt and Brennan (1989) have observed that, because reliability is easier to determine using mathematical formulae, it has become the main preoccupation of many measurement specialists for the greater part of the twentieth century. Validity, which is more important but can be studied only by considering a number of related issues, has lagged behind and has only begun to receive meaningful attention during the latter half of the twentieth century. The view has also prevailed that reliability should really be merged into the study of validity. This view has been

aptly propounded by Feldt and Brennan (1989): "No body of reliability data, regardless of the elegance of the methods used to analyze it, is worth very much if the measure to which it applies is irrelevant or redundant" (p. 143).

However, as stated in the *Standards for Educational and Psychological Testing* (1985), basic to the proper evaluation of any test is the identification of major sources of measurement error, the size of the errors that result between the pairs of scores under particular circumstances, and the generalisability of results across items, forms, raters, administrations, and other measurement aspects.

If the existing literature on psychometrics appears more weighted in favour of reliability, it is because reliability is so uniquely linked with measurement error. As Feldt and Brennan (1989) observed, reasonable assumptions and deductions about measurement error can be easily stated in mathematical terms. Secondly, part of the explanation for a greater array of literature on reliability derives from the fact that an investigation of reliability is possible on the basis of test data only, without recourse to any external data. Finally, there is greater importance of subjective judgment associated with validity compared to reliability. Questions about the adequacy of criteria, the defensibility of definitions of human traits, the appropriateness of test content, the clarity of the boundaries of behavioural domains, and the implications of data for all these matters inevitably involve the exercise of judgment. "A game played by subjective, rather than mathematical, rules may be harder to play well, more prone to professional controversy, and attract fewer players" (Feldt & Brennan, 1989, p. 143).

Issues Pertaining to Validity

Cronbach (1971) in a classic article, *Test Validation*, described test validation as a process in which evidence is collected by the developer of a test to support the types of inferences that might be appropriately drawn from test scores. Embedded in this

view of validity is the idea that the emphasis of validity is not on the instrument itself; rather, the emphasis is on the interpretation of the scores yielded by a test.

Later Cronbach (1988) addressed the changing nature of validity by describing it as a concept that has to be viewed from varying perspectives, some of which were captured by other researchers, as indicated below:

> The term validity, when applied to a test, refers to the precision with which the test measures some cognitive ability. (Ebel & Frisbie, 1986, p. 89)
>
> The validity of a test concerns what the test measures and how well it does so. (Anastasi, 1988, p. 139)
>
> Validity is an overall evaluative judgment, founded on empirical evidence and theoretical rationales, of the adequacy and appropriateness of inferences and actions based on test scores. As such, validity is an inductive summary of both the adequacy of existing evidence for and the appropriateness of potential consequences of test interpretation and use. (Messick, 1988, pp. 33-34)
>
> Validity can best be defined as the extent to which certain inferences can be made from test scores or other measurement. (Mehrens & Lehmann, 1987, p. 74)
>
> Validity is defined as the extent to which measurements are useful in making decisions relevant to a given purpose. (Sax, 1980, p. 289)

Concepts of validity have shifted over time. The 1974 *Standards for Educational and Psychological Tests* (AERA, APA, & NCME, 1974) defined validity in three interrelated ways: content, construct, and criterion validity. By 1985, however, the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1985) discussed validity as a unitary concept requiring multiple types of evidence: construct-related evidence, content-related evidence, and criterion-related evidence (Moss, 1992).

Role and Meaning of Validity in Educational Measurement

> Validity is the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. Test validation is the process of accumulating evidence to support such inferences. A variety of

> inferences might be made from scores produced by a given test, and there are many ways of accumulating evidence to support a particular inference. Validity, however, is a unitary concept. Although evidence might be accumulated in many ways, validity always refers to the degree to which that evidence supports the inferences that are made from the scores. The inferences regarding specific uses of a test are validated, not the test itself. (*Standards for Educational and Psychological Testing*, 1985, p. 9)

Traditionally, the numerous ways of collecting evidence for validity have been grouped into categories called content-related, criterion-related, and construct-related evidence for validity. These categories are a convenient way of treating validity, and so are the subcategories of predictive and concurrent validity within criterion-related evidence. The use of these categories does not imply that there are different types of validity or that one approach is the best for each specific inference or test use. Indeed, rigorous distinctions among the categories are not possible because evidence identified under one category is usually relevant under another. Generally, the more sources there are for validity evidence, the better. However, the quality of evidence is of paramount importance. In quite a number of cases a single line of solid evidence is often preferable to numerous lines of evidence of dubious quality.

According to Ary et al. (1990), Hopkins et al. (1990), Mehrens and Lehmann (1991), Sax (1989), and the *Standards for Educational and Psychological Testing* (1985), content-related evidence, criterion-related evidence, and construct-related evidence for validity are only three different ways of looking at the same thing, as the following short descriptions of each will show.

Construct-Related Evidence

Construct-related evidence for validity is the most important category of validity evidence. In fact, it is the unifying concept, and that which we call content- and criterion-related validity could easily be included within the perspectives of trying to clarify constructs through content definitions and domains and through correlations of the construct with other variables (Messick, 1988). Construct-related evidence studies

combine both a logical and an empirical approach. A measure of a particular construct should be as independent as possible of measures of unrelated constructs. If, for instance, we develop a test designed to measure arithmetic problem-solving skills and we find that the scores obtained are highly correlated with scores of reading tests, we would conclude that we have developed another test of reading rather than an arithmetic problem-solving test per se. However, a logical analysis of test content would reveal that the test is measuring arithmetic problem-solving and reading ability.

Construct-related evidence studies involve probing a number of sources of evidence such as correlations, content analysis, and protocols. When a new test of a construct correlates highly with other tests which measure the same construct, then convergent validity evidence for the new test has to be found. On the other hand, low correlations between a measure of a particular construct and measures of other constructs provide evidence of discriminant validity for the measure in question. The presence of discriminant validity also implies divergent validity. In short, tests of similar concepts (convergent validity evidence) should always be highly correlated, whereas measures of dissimilar constructs should always be lowly correlated (discriminant or divergent validity). Evidence from content analysis and individual protocols can be used to confirm the extent to which a particular psychological construct is reflected in the analyses of results.

Content-Related Evidence

Content-related evidence shows the extent to which a sample of items on a test is representative of some defined universe or domain of content. The evidence is gathered through a careful and critical examination by expert judges of the test's content to determine the relationship between the test and the defined universe.

Criterion-Related Evidence

Criterion-related evidence shows the extent to which scores on a measuring instrument are related to an independent external variable (criterion) believed to measure directly the behaviour or characteristic in question. The overarching principle here is that the criterion chosen should be relevant, reliable, and free from bias, especially criterion contamination which occurs when an individual's score on the criterion is influenced by the scorer's knowledge of the subject's predictor score. In the study of criterion-related evidence, a distinction is often made between predictive validity and concurrent validity. Predictive validity evidence involves the correlation between the test scores and the criterion scores obtained at a future time; for example, Advanced Level performance (the predictor) and performance at university (the criterion), or the relationship between the Zimbabwe Junior Certificate (ZJC) performance (the predictor) and performance at the Ordinary Level (the criterion). Concurrent validity studies gather information about the correlation between test scores and a criterion measure obtained at the same time. For example, we might correlate grades obtained by candidates in the ZJC examination in particular subjects with grades given by teachers in the same subjects—the so-called internal versus external assessment components would provide evidence of concurrent validity. Predictive validity evidence is generally preferred in the validation of selection tests in education or industry, whereas concurrent validity evidence is preferred for achievement tests, tests designed to measure constructs, or tests used for certification or diagnosis (Cook, 1991).

Validity Generalisation

According to the *Standards for Educational and Psychological Testing* (1985):

> Validity generalisation refers to the extent to which there are similarities between one situation and another that could influence researchers to transport conclusions made in one situation to the other. In studies on

> validity generalisation, the major aspects that have to be considered are (a) differences in the way the predictor construct is measured, (b) the type of curriculum followed, (c) the type of criterion measure chosen, (d) the type of test takers, and (e) the time period in which the study is conducted. In any particular study of validity any number of these features might vary, and a major objective of the study is to determine whether variation in these factors affects the generalisation of validity evidence. (p. 12)

Differential Prediction

Differential prediction is a broad concept that includes the possibility that different prediction equations can be obtained for different demographic groups, for groups that differ in their prior experiences, or for groups that receive different treatment or are involved in different instructional programmes. The term *treatment* is intended to convey the concept of not only the various forms of intervention in which subjects might be involved, but also the way in which the tests are administered, such as by computer. In the case of the testing scene in Zimbabwe, an interesting situation might arise in differential prediction for boys and girls, on the one hand, or former white-only schools versus former African-only schools, on the other. Consideration should be given to different criterion predictions that result from different groups of test takers in order to achieve fairness where important decisions such as selection or classification are to be made.

Validity and Test-Wiseness

According to Messick (1989), *test-wiseness* refers to a situation in which partial knowledge of the content is combined with the notion of test-wise-susceptible items to improve a subject's test score. The principle of test-wiseness is often used to explain the construct-irrelevant easiness of a number of test items. According to Benson (1988) and Sarnacki (1989), test-wiseness is a cognitive ability or a set of skills a test taker can use to improve a test score irrespective of the content area. The most

popular definition of test-wiseness that has often been adopted in many research studies has been that proposed by Millman, Bishop, and Ebel (1965): "a subject's capacity to utilize characteristics and formats of the test and/or test-taking situation to receive a high score" (p. 707). Taken together, these definitions suggest that test-wiseness encompasses both the method of measurement (items that provide test-wise cues) and characteristics of the test taker (cognitive abilities or set of skills that examinees could employ in any testing situation regardless of the content measured).

If a test taker possesses test-wiseness and if the examination contains susceptible items, then the combination of these two factors could result in an improved score; in contrast; a student low in test-wiseness would tend to be penalised every time he or she takes a test that includes test-wise components. Thus, a potential validity problem exists when one attempts to interpret the meaning of the test score (Rogers & Bateson, 1991).

According to Rogers and Bateson (1991), the following elements of test-wiseness cues were most commonly found in school-leaving examinations:

1. Three deductive-reasoning strategies:
   - ID1 - Eliminate options known to be incorrect (absurd distractors).
   - ID2 - Choose neither or both of two options which imply the correctness of each other (similarities).
   - ID3 - Choose neither or one of two options, one of which, if correct, would imply the incorrectness of the other (opposites).
2. One cue-using strategy:
   - IIB4 - Recognise and use similarities between the stem and options (cued options).

In a study carried out by Rogers and Bateson (1991) in British Columbia, the percentage of test-wise-susceptible, four-option multiple-choice items varied from 43% to 80% across the provincial school-leaving examinations in English, algebra,

geography, history, biology, and chemistry. In a replication of this study, similar percentages (70.0% to 87.5%) were found across five school-leaving examinations in Alberta (Rogers & Wilson, 1993).

Gender Issues in Performance

The list of research studies published on gender issues in performance is a large and increasing one. In their pioneering work, Maccoby and Jacklin (1974) reviewed the enormous literature prior to 1974 on psychological gender differences and concluded that verbal ability, quantitative ability, and visual-spatial ability reflect cognitive gender differences. In particular, they reported that, while girls have greater verbal ability than boys, boys have better visual-spatial ability and mathematical ability than girls. Sherman (1978) re-reviewed the evidence on cognitive gender differences and observed that gender differences in all the abilities identified by Maccoby and Jacklin in 1974 were very small, varying from .24 standard deviations for verbal ability to .45 for visual-spatial ability. In addition, the proportion of variance accounted for by gender differences was also found to be very small, ranging from .01 for verbal ability to .04 for visual-spatial ability. Similar findings were confirmed by Hyde (1981). Feingold (1988) was able to demonstrate that cognitive gender differences which had been identified and confirmed to be small were actually disappearing in all subject areas except high school mathematics. These declining gender differences were also confirmed by Jacklin (1989), who concluded:

> In summary, tests of intellectual ability have differentiated girls and boys less and less over the last decades. The only exception to this trend is at the highest end of the mathematics continuum, where the ratio of boys outscoring girls has remained constant over the years. (p. 128)

Jacklin actually hoped that there would be less and less research on this topic in future, given the trend of declining gender differences. Whereas this might have been the case in North America, the issue of gender in performance is far from being

exhausted, especially with reversed trends now being experienced in the United Kingdom and Zimbabwe (Elwood, 1994; Gipps, 1994; Nhandara, 1993; Stobart et al., 1992).

As recently as 1992 Harris and Pickle (1992) made an impassioned plea for gender equity. In their argument, Harris and Pickle observed that girls' achievement was still problematic although girls started out with an academic advantage because they matured earlier than boys. They further observed that, whereas boys generally judge themselves by what they are able to do, girls generally portray their worth in terms of their physical appearance: "Thus girls exhibit constrained views of their potential and their 'place' in society and much less confidence in their abilities—particularly in math and science" (p. 12). In trying to provide evidence for their observations, Harris and Pickle cited a number of factors found in current research (Feingold, 1988; Hyde, 1981; Jacklin, 1989; Maccoby & Jacklin, 1974; Sherman, 1978) which supported them. In particular, Harris and Pickle mentioned that (a) boys receive more teacher attention, praise, and blame and are asked more higher-level questions than are girls; (b) the curriculum was derived more from the productive rather than the reproductive world; (c) classroom learning methods emphasize ritual opposition such as argument and challenge; and (d) it is a common practice for students to reconstruct "history in terms of wars, rulers and territory, often depreciating the emotional work that is necessary to sustain harmony in human relationships" (p. 12) as some of the major factors that perpetuate gender differences in performance.

Using the Scholastic Aptitude Test (SAT), Harris and Carlton (1993) confirmed gender differences on mathematics items. However, they observed that there are identifiable patterns of gender differences in how male and female students arrive at their total scores. For example, male students perform relatively better on geometry

and geometry/arithmetic items, as compared with matched female students who perform relatively better on miscellaneous and arithmetic/algebra items.

In a detailed survey of gender bias in the General Certificate of Secondary Education (GCSE) introduced in the United Kingdom in 1988 for the 16+ age group, Stobart, Elwood, and Quinlan (1992) noted some important and interesting observations which, given the background that the examination system in Zimbabwe has tended to follow closely that of the British, have a direct implication for Zimbabwe. Basing their observations on the GCSE examinations spanning the period 1988 to 1990, Stobart et al. observed that there were significant gender differences in both the entry patterns and outcomes. In English girls achieved better grades in most subjects except mathematics. During the same period entry patterns seemed to be guided by and reflected the level of subject difficulty and the confidence with which individuals viewed a particular subject. Such a situation seems to be confirmed by the entry patterns shown in Table 2, where girls made up only 28.4% of the physics entry and 44.5% of the chemistry entry, whilst constituting 62.8% of the biology entry. Other subjects which attracted a relatively high proportion of girls in comparison to boys were English literature and French, with an average entry of 59.0% and 53.4%, respectively, for the 1988-90 period. It could be safely concluded then that boys generally avoid the languages (in keeping with research findings) if they have a choice. Where they are forced to take a subject, as in English, their performance is lower than that of girls. As shown in Table 3, for the period under study, English and mathematics (the two compulsory subjects) produced very different outcomes, with 54.6% of the girls achieving Grades A-C compared with 41.5% of the boys. The pattern was reversed in mathematics, where 34.6% of the girls gained Grades A-C, whereas 38.9% of the boys did so. The cohort taking the two subjects was pretty much the same.

Table 2

All GCSE Groups' Entries 1988-90

Subject	Male % of entry 1988	Male % of entry 1989	Male % of entry 1990	Female % of entry 1988	Female % of entry 1989	Female % of entry 1990
Biology	36.9	37.5	38.9	63.1	62.5	61.1
Chemistry	55.6	55.3	55.1	44.4	44.7	44.9
Economics	60.5	63.8	61.4	39.5	36.2	38.6
English	49.0	49.4	49.7	51.0	50.6	50.3
English literature	44.6	49.4	45.9	55.4	50.6	54.1
French	40.2	41.0	41.8	59.8	59.0	58.2
Geography	58.7	58.3	57.9	41.3	41.7	42.1
History	48.8	49.1	48.8	51.2	50.9	51.2
Mathematics (with coursework)*	48.2	48.2	48.3	51.8	51.8	51.7
Mathematics (no coursework)*	47.8	48.9	48.6	52.2	51.1	51.4
Physics	72.5	71.6	70.8	27.5	28.4	29.2

* Coursework optional in GCSE mathematics until 1991.
Source: 1988, 1989, & 1990 *Inter-Group Statistics.* (Stobart et al., 1992, p. 265)

Stobart et al. (1992) also examined some of the important factors that contribute toward gender differences in examination performance. Amongst these are individual prior experiences and expectations and types of assessment techniques used.

Within the languages, differences could come about as a result of the different reading and writing styles. Girls read a wide range of books, including fiction, whereas boys stick to nonfiction material, especially technical manuals and information sources. During writing, girls tend to use extended reflective composition, whereas boys prefer to provide episodic, factual, and commentative detail for the same task. In view of the stronger emphasis which examiners now place on linguistic abilities in writing compositions, boys are at a disadvantage owing to their reading and writing styles (Elwood, 1994; Stobart et al., 1992).

Table 3

All GCSE Groups' Percentage Grades A-C, 1988-90

Subject	Male (M) % A-C			Female (F) % A-C			Difference (F-M) (%)			Mean difference 1988-90
	1988	1989	1990	1988	1989	1990	1988	1989	1990	
Biology	48.7	51.8	53.9	41.0	44.4	47.0	- 7.7	- 7.4	- 6.9	- 7.3
Chemistry	50.9	53.3	55.2	47.1	49.7	52.7	- 3.8	- 3.6	- 2.5	- 3.3
Economics	47.8	53.5	53.9	45.3	54.5	52.0	- 2.5	1.0	- 1.9	- 1.1
English	36.8	41.6	46.0	50.5	55.2	58.0	13.7	13.6	12.0	13.1
English literature	44.8	41.9	50.7	56.5	55.4	63.1	11.7	13.5	12.4	12.5
French	46.9	49.9	45.6	52.1	55.3	51.9	5.2	5.4	6.3	5.6
Geography	39.0	43.6	44.0	45.1	49.0	51.0	6.1	5.4	7.0	6.2
History	41.0	45.5	46.2	46.7	51.1	52.3	5.7	5.6	6.1	5.8
Mathematics (with coursework)*	36.4	41.3	42.9	32.4	36.8	38.0	- 4.0	- 4.5	- 4.9	- 4.5
Mathematics (no coursework)*	42.5	41.9	43.2	32.8	35.6	37.9	- 9.7	- 6.3	- 5.3	- 7.1
Physics	43.5	48.6	52.3	47.5	52.7	56.3	4.0	4.1	4.0	4.0

* Coursework optional in GCSE mathematics until 1991.
Source: 1988, 1989, & 1990 GCSE *Inter-Group Statistics.*

(Stobart et al., 1992, p. 267)

What a student can do with a subject after school greatly influences interest and effort in that subject and hence the grade obtained at the end (Gipps, 1994; Murphy, 1994). It could be conceded (wrongly, for that matter) that the view still prevails in many societies that girls do not need mathematics as much as boys do in the world after school; hence their poor performance in comparison to boys. Another source of evidence for the argument above is Table 4, which shows that girls entered basic or foundation mathematics, whereas boys tended to enter higher mathematics.

According to Stobart et al. (1992), the type of assessment technique used directly influences the outcome. Over the years it has been demonstrated that, generally, boys do much better on multiple-choice examinations because they tend to settle on one out of many options with greater confidence (even if the option is wrong) than do girls, who are adversely affected by the desire to weigh the relative

Table 4

GCSE Mathematics Entry Patterns 1988-90

	1988		1989		1990	
Level	Male	Female	Male	Female	Male	Female
Foundation (E-G)	78,974	85,562	69,603	76,127	51,041	54,505
% of total	48.0	52.0	47.8	52.2	48.4	51.6
Intermediate (C-F)	144,209	162,721	142.448	166,298	118,970	141,558
% of total	47.0	53.0	46.1	53.9	45.7	54.3
Higher (A-C)	97,680	87,203	89,529	78,754	71,335	64,802
% of total	52.8	47.2	53.2	45.8	52.4	47.6
Total entry	320,863	335,576	301,580	321,179	241,346	260,865
% of total	48.9	51.1	48.4	51.6	48.1	51.9

Source: 1988, 1989, & 1990 *Inter-Group Statistics.* (Stobart et al., 1992, p. 269)

rightness/wrongness of a number of options. In contrast, boys have more difficulty expressing themselves in written English than girls do.

The situation described above could be the main reason why in Zimbabwe during the 1970s more boys than girls obtained Grades A-C in the Cambridge Ordinary Level English Language Examination, with its entire Paper 2, worth 50% of the total marks, composed of multiple-choice items. When this component was discontinued, there was a sudden significant drop in the number of superior grades for the male population.

As reported in Table 4, the universal introduction of a coursework component in every GCSE subject has, with the exception of economics, led to significantly improved grades for girls, even in those subjects where boys had previously outscored girls. A classic case in point is that of history, where the trend was reversed (see Table 5). In 1985 boys outscored girls at Grades A-C by 3.1%. The trend was reversed in 1988, when girls outscored boys by 5.7%. A similar trend was also observed in geography, where girls outscored boys by 6.1% after an almost equal performance in 1985. Coursework, wherever it has been used as an assessment component with a substantial weight (e.g., 30-50% of total subject grades), has benefitted girls more than it has boys because girls are characterised by consistent hard work throughout the course period, whereas boys prefer to work for the examination.

Current Thinking on Validity

Lumsden (1976) observed that most measurement theorists have fled from validity as a central point in educational measurement largely because the concept is too difficult for simple-model makers. Unlike reliability, validity does not lend itself to neat mathematical models from which correlation indices can be obtained. Test validation is "a process of coming to some understanding of the meaning of test

Table 5

Comparison of Gender Differences (Female-Male) for 1985 GCE and for 1988 GCSE

Subject	1988 GCSE* grades A-C (%)	1985 O-level*† grades A-C (%)
Biology	- 7.7	- 6.2
Chemistry	- 3.8	- 2.5
Economics	- 2.5	- 4.0
English	13.7	3.7
French	5.2	2.5
Geography	6.1	- 0.1
History	5.7	- 3.1
Mathematics	- 4.0	- 6.5

* Includes figures for England, Wales, and Northern Ireland.
† Source: O-level *Inter-Board Statistics*, June 1985.
‡ Source: DES *Statistics of Leavers CSE & GCE, England 1985.*

(Stobart et al., 1992, p. 273)

scores for various applications of the test procedure" (p. 268). Seen in this perspective, validity is essentially a judgment call. Cronbach (1988) has succinctly captured the shift in emphasis on validation procedures from being exclusively the province of the test developer to being a concern for everyone:

> Validation was once a priestly mystery, a ritual performed behind the scenes, with the professional elite as witness and judge. Today, it is a public spectacle combining the attractions of chess and mud wrestling. Disputes about the appropriateness of tests impose a large responsibility on validators. (p. 3)

In many countries public pressure and legislation have resulted in assessment objectives (specification tables) and assessment schemes, including test samples, being made available to students and their teachers in a new move designed to open up the testing process to public scrutiny. Such activities encouraged Cronbach (1988) to argue for a broadened view of validity in an effort to apply what he called the logic of evaluation argument to the validation of test interpretations or test use and to

emphasise validity argument rather than validation research. Such a perspective has also emphasised the need to understand the context of test use, as well as to understand what generated test scores.

With respect to test use and test scores, Mehrens (1984) observed that "not all achievement tests measure the same thing. No achievement test measures everything" (p. 9). Mehrens continued and elaborated on his argument:

> Actually, it is good that not all tests measure exactly the same thing. It allows those who wish to infer to slightly different domains to choose different tests. No achievement test measures all the instructional objectives of a local school. That is, there will be a taught but not tested mismatch. This is not a problem as long as no one is foolish enough to infer that tests do measure everything. I know no one that foolish. (p. 14)

In carrying out validity studies, the central focus has shifted and is now centred on the concept of construct validity. All the other sources of validity evidence, such as content-related and criterion-related, are viewed as extensions or elaborations of construct validity. The numerous ways through which validity is now being studied can be summarised by the following six approaches:

1. a judgmental and logical analysis of content relevance and representativeness of a given test, such as the one indicated by the specification table;

2. carrying out correlational studies in order to examine the structure and covariance components with respect to item or task consistencies and relating test scores to other variables such as trait validity;

3. carrying out protocol analysis in order to probe the thinking processes and test behaviours underlying item responses and task performances;

4. investigating differences in test processes and structures over time or across groups and settings, as in longitudinal and cross-sectional validity studies;

5. examining score changes in response to instructional intervention or experimental manipulations of test content and test conditions; and

6. appraising the value implications and social consequences of interpreting and using test scores in particular ways, such as the undesired social side effects when all those who have failed a particular test are often regarded as "academic failures" or morons.

According to Smith and Glass (1987), the six approaches outlined above are all part of the overarching category of construct validity intended to address the following four basic questions:

1. Was there logical consistency between the content of the test items, observation schedule, rating scale, and so on, and the definition of the construct? Such construct validity studies are primarily aimed at ensuring that the test data is interpreted as the evaluator intended.

2. Has an empirical connection been established between the performance on the chosen indicator (test) and some other indicator that purports to measure the same construct? Such concurrent and predictive validity studies illustrate the extent to which a particular test correlates with similar current measures or similar future assessments.

3. Has an empirical connection been established between the indicator (test) and other indicators of the same construct that might use different methods? Such convergent validity studies should yield high coefficients for the two assessment to be considered as measuring the same construct.

4. "Have successful efforts been made to establish that the indicator of a construct does not correlate with indicators of other constructs that are theoretically unrelated to the test?" (Smith & Glass, 1987, p. 109). According to Campbell and Fiske (1959), when a test correlates too highly with other tests from which it is intended to differ, its validity becomes suspect. A low correlation between a test and other indicators that are theoretically unrelated to it is evidence of the test's discriminant validity.

Messick's (1989) study on validity has often been taken as the state-of-the-art explanation of construct validity. Much of the information is generally correct and can be viewed as a synthesis of the work undertaken by Cronbach and Meehl (1955), Loevinger (1956, 1957), Ebel (1961), Cronbach and Gleser (1965), and Cronbach (1971, 1988). Maguire, Hattie, & Haig (1993) have strongly challenged some of Messick's conclusions. They have observed that

> the unifying structure of Messick's chapter is the two-dimensional Facets of Validity Table [see Figure 2] in which the "Functions or outcomes of testing" (split into test interpretation and test use) are crossed with the "sources of justification" (divided into evidential bases and consequential bases) to produce a fourfold table: 1. the evidential basis of test interpretation—construct validity, 2. the evidential basis of test use—construct validity incorporating relevance and utility, 3. the consequential basis of test interpretation—construct validity and value implications, and 4. the consequential basis of test use—construct validity, relevance, utility and the value and social consequences. (p. 4)

	Test Interpretation	Test Use
Evidential Basis	Construct Validity	Construct Validity + Relevance/Utility
Consequential Basis	Value Implications	Social Consequences

Figure 2. Facets of test validity (Messick, 1988, p. 42).

There seems to be an obvious imbalance in Messick's (1989) approach to validity with respect to achievement testing vis-à-vis large-scale commercial testing. Maguire et al. (1993) have observed that Messick's argument for consequential validity as a central validity failed because of three underlying reasons:

1. There is an overemphasis on the role of the developer of large-scale standardised tests at the expense of achievement tests, which contribute the largest proportion of educational tests in the world today.

2. Too much emphasis is placed on large-scale systematic and planned assessments of consequential validity as if all testing today were some kind of industry just like a pharmaceutical centre whose products are judged solely from the effectiveness associated with their use. Clearly, the bulk of educational tests today do not belong to this category, which emphasises approved and disapproved uses.

3. Overemphasis on consequential validity favours test-score use rather than test development. The stress on consequential validity seems to be a reaction to the increased number of court cases surrounding tests and testing practice, specially of the large-scale standardised, commercial type.

The view held by Maguire et al. (1993) is not that the consequences of test use are unimportant.

> Rather, it is our contention that a concern with consequences should be moved out from the umbrella of construct validity and into the arena of ethical guidelines such as the *Principles of Fair Assessment Practices for Education in Canada (1993)* and informed debate. The *Principles* deal not only with items contained under Messick's consequential validity, but speak of follow-up and redress as well. They place ethical test use properly in the arena of professional responsibility and encourage an atmosphere of openness and questioning. (p. 10)

Maguire et al. (1993) have also raised important issues which all have a direct bearing on construct validity. In particular, they have questioned the widespread use of scoring models which are based on the classical test theory and result in teachers and testing organisations simply adding up the item scores to arrive at a total score, with no regard at all to how examinees arrived at their different total test scores. They have rightly observed that many scoring systems have never been validated. They viewed tests and scoring procedures as "mediating empirical operations" (p. 19) to make constructs into numbers, and in that process "it is not obvious that the

common practice of administering large numbers of items and computing number correct, or weighted aggregates, retain the representative link between numbers and constructs" (p. 19).

Maguire et al. (1993) have also pointed out the shortcomings deriving from the table of specifications, where content is regarded as being static and where issues such as the context and relative position of content and the interactive nature of cognition and content have tended to be ignored. In addition, factor analysis and structural modelling, which have been widely used as test validation processes, have their limitations in that they are based on metric assumptions of the nature of relationships among the constructs that they help us to investigate. Both also encourage sloppy thinking about the nature of causal relationships.

Whatever else might have been done in the past and will continue to be done in the future in the study of construct validity, one cannot be criticised for concluding that the notion of validity is not a case for considering all facets, but a question of the degree to which some facets enhanced the interpretation of test scores. Gronlund (1993) has, indeed, captured the essence of validity in his six points on validity:

1. Validity is inferred from available evidence (not measured).
2. Validity depends on many different types of evidence.
3. Validity is expressed by degree (high, moderate, low).
4. Validity is specific to a particular use.
5. Validity refers to the inferences drawn, not the test itself.
6. Validity is a unitary concept. (p. 160)

Chapter 3

Research Methodology and Procedures

The primary objective of this study was to develop a psychometric model upon which educational tests in Zimbabwe could be both constructed and analysed. Specific methods and procedures employed during the study are described under relevant headings in this chapter. The first section describes the instruments used, followed by a description of the sample in the second section. The third section contains information on the administration of the instruments, followed by the scoring and preparation of data in the fourth section. The last section describes the procedures used to analyse the data. Geography was chosen as the subject in which to develop a psychometric model for testing in Zimbabwe for two basic reasons. In the first place it was the subject area of the researcher, and secondly, it was viewed as a bridging subject between the arts and sciences.

Instruments

The test administered to the examinees was the 1987 ZJC Geography Specimen Paper 1 (Multiple-Choice), developed by the National Geography Panel. The 24-member National Panel was constituted as follows: 18 (2x9) representatives from each of the country's nine education regions, 2 representatives from the Curriculum Development Unit, 1 representative from the Standards Control Unit (Schools Inspectorate Coordinating Unit headquartered in the Ministry of Education and Culture), 2 representatives from the secondary teacher-training colleges, and 1 representative from the University of Zimbabwe. All panelists were university geography graduates, and, out of the total of 24, only 2 panelists were females—the rest were males.

With one exception, there was no table of specifications at the time the specimen paper was developed by the panel of item writers. The one exception was

the topic area of mapwork, which was allocated 12 of the 50 items to be included in the paper.

In order to develop the specimen paper, panelists were requested to generate items in the ZJC syllabus topic area of their interest. Once generated, such items were submitted to the Curriculum Development Unit for collation before a full-panel item-review meeting. With regard to the ZJC specimen Paper 1, three panel item-review meetings were held before the specimen paper could be compiled. The final selection of items and the construction of the paper were left to be carried out by the Curriculum Development Unit.

The test itself, consisting of 50 items and its subdivisions into four basic content areas of mapwork, physical geography economic geography, and population and settlement, is very similar to the actual geography examination given at this level. In terms of the overall subject grade, the multiple-choice paper accounted for 50% of the total marks, the other 50% being contributed by the structured and free-response paper. A copy of the paper is provided in Appendix B.

Following development of the ZJC Geography Specimen Paper 1, it was planned to conduct a field test prior to its actual use. However, there were insufficient funds to carry out this field test. Therefore, this research can be seen as taking the place of the initially planned field test. This field test took place during the period June to August 1993. This period was considered ideal because the majority of candidates would have completed their ZJC syllabuses. Mentally, the candidates were considered to be in an examination frame because they were writing their mock ZJC examinations during the same period.

The Sample of Examinees

The sample of examinees for the present study consisted of 909 students from 16 (21.6% of total) high schools in the Harare Region (one of the nine education regions into which Zimbabwe is divided). Of the total, 495 were boys and 414 were girls. The group of examinees was a purposive sample which was deliberately selected to include the full range of examinee ability (high, average, low) at this level. The actual choice of examinees was carried out on the basis of whole classes selected with the help of geography teachers at each of the schools.

Apart from two exceptions (rural council and mine/farm high schools), the 16 high schools were representative of the types of schools found in Zimbabwe (see Table 6, which shows the types of high schools that were included in the test sample).

The students were in their second year of high school and preparing for the ZJC examination, which they were scheduled to write during October/November 1993 Examinations. Geography was included as one of the subjects in the set of papers to be written.

Administration of the Test

At all 16 schools, the test was personally administered by the researcher. At schools where examinee numbers required the use of two rooms, the researcher's wife and five geography teachers assisted with invigilation of the other group. The examinees wrote the test in their classrooms and, where possible, during the periods normally timetabled for geography. Such an arrangement was made in order to cause as little disruption as possible to the school day. All examinees were allowed a maximum of 1-3/4 hours to write the test.

Table 6

Types of Schools in the Harare Region (1993) Included in the Test Sample

Name	Type	Number of subjects		
		Girls	Boys	Total
Epworth	7	13	28	(41)
Morgan	1	29	36	(65)
Mbare	2	19	23	(42)
Churchill	1	—	75	(75)
St. Mary's	3	16	23	(39)
Nyatsime	3	19	19	(38)
Zengeza	2	43	41	(84)
Prince Edward	1	—	68	(68)
Girls' High	1	58	—	(58)
Mt. Pleasant	1	15	21	(36)
Mufakose 1	2	35	40	(75)
Eaglesvale	4	12	14	(26)
Glen Norah 1	2	37	36	(73)
Glen View 2	2	37	36	(73)
Lord Malvern	1	45	35	(80)
Dominican Convent	4	36	—	(36)
		414	495	909

Note:
1: Government (former whites, coloured, Asian-only school)
2: Government (former Africans-only school)
3: Private (former Africans-only—church related)
4: Private (former whites, coloured, Asian—non-church related)
7: Private (former African = non-church related)
* Types 5 and 6 are located in other regions where rural councils and local authorities such as mines and farms are in control.

Scoring and Data Preparation

The answer sheets (optical mark readers) were coded according to school, examinee identification number, and gender before being brought back to Canada for scoring at the University of Alberta optical mark scoring unit. A total of four computer programmes was used to analyse the data; namely, (a) the Laboratory of Educational Research Test Analysis Package (LERTAP) (developed by Larry Richard Nelson in the mid 1970s), to conduct a classical test score theory item analysis and compute estimates of intended consistency for each of the four subtests and the total test and Cronbach's alpha for the composite test; (b) Item Anal, a local University of Alberta programme developed by Dan Precht of Network Systems for scoring and classical test score theory item analysis; (c) SPSSx (a statistical package), used for computing t-tests for two independent samples; and (d) BICAL (One Parameter Rasch Model), used to conduct item response theory analysis.

Procedures for the Analysis of Results

Internal consistency indices. In order to investigate the reliability of the test, the 50 items were subjected to internal consistency for the total test, which was computed using Hoyt's (1941) analysis of variance procedure. As stated in Chapter 2, given that the items were dichotomously scored, this procedure is equivalent to the procedures used to compute Cronbach's alpha and K-R20. In addition, the internal consistency was computed for each of the four subtests; namely, (a) mapwork, (b) physical geography, (c) economic geography, and (d) population and settlement. The various subtest internal consistency coefficients were used to illustrate the extent to which the items included in each subtest were internally consistent, and the concept of the impact of test length and other factors such as examinee ability and item difficulty on reliability. Lastly, Cronbach's composite alpha was computed using the subtests as the unit of analysis. This index provided an

indication of the degree to which the subtests yielded consistent scores. A high value would suggest that the subtests were essentially homogeneous. Conversely, a low value would suggest that the subtests had little overlap and were measuring different aspects of achievement in geography. These analyses were completed using the LERTAP computer programme.

Item Analysis Using Classical Test Score Model

At the item level, LERTAP and Item Anal yield two indices used to assess the functioning of an item, namely, the p-value (the proportion of examinees who got the correct answer) and the item discrimination (measure of how effectively an item differentiated between examinees with the highest and lowest scores on the total test). Item discrimination was reflected by three indices, namely, the point-biserial correlation (rpbis), the biserial correlation (rbis), and the D statistic. These three indices of item discrimination are discussed below in some detail. LERTAP also gives option means which can be used to assess discrimination.

Point-biserial and biserial correlations. According to Crocker and Algina (1986) and Osterlind (1989), biserial and point-biserial correlations are closely related because both give an indication of the relationship of the performance on an item and performance on the total test for the high and low achievers. The distinction between these two measures lies in the assumptions made. Whereas the point-biserial statistic assumes that the item score is a true dichotomy (scored right or wrong), the biserial correlation assumes that both the item score and total test score are inherently continuous (item score is treated as an artificial or forced dichotomy).

Point-biserial correlation. Under the point-biserial correlation, the genuine dichotomy (item score) is assigned numerical values of 1 or 0, and these values are correlated with the values on the continuous variable (total test score). The point-biserial correlation is considered the most effective indicator in examining the relative

performance of an item between the high- and low-ability groups, and is therefore the most frequently used. In general, items with higher point-biserial correlations for the correct option are more highly discriminating. Conversely, items with low point-biserial correlations are less discriminating. As a general rule, items with negative point-biserial correlations for the correct option are usually very difficult and reflect a great deal of random guessing from both the high- and low-ability groups. Such items are either dropped from a test or improved before being included in a test. It would also be less desirable to have items that are missed by many high-scoring examinees but are answered correctly by low-scoring examinees (see Figures 3-6). For options (foils) to be considered effective, they must attract an equal number of examinees, because if they do not do so they will be considered flawed. Further, their point-biserials should be negative (less than -.10).

The point-biserial correlation, which is the popular statistic in conventional item analysis, presents two problems. The first problem is that the single item being analysed has itself contributed to the total test score or the ability measure. This problem is considered to have minimal effect, especially when the number of items is relatively large; say, 25 or more. Where the number of items is very small—say, below 25—the problem may be corrected for attenuation using the formula

$$\rho i(x\text{-}i) = \frac{\rho xi\sigma x - \sigma i}{\sqrt{\sigma^2 i + \sigma^2 x - 2\rho xi\sigma x\sigma i}}$$

where $\rho i(x\text{-}i)$ is the correlation between an item score and the total score with that item removed, and σx and σi are the total and item standard deviations, respectively.

Otherwise, the ordinary formula for point-biserial correlation is

$$\rho pbis = \frac{(\underline{Uc} - \underline{Ux})}{\sigma x} \sqrt{p/q}$$

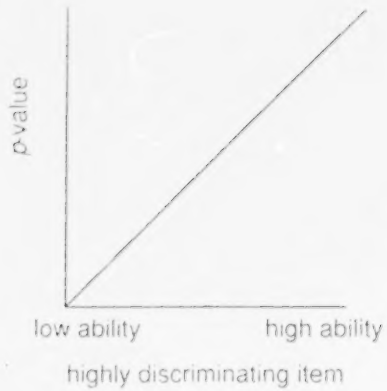

Figure 3. Characteristics of a highly discriminating test item (Osterlind, 1989, p. 284).

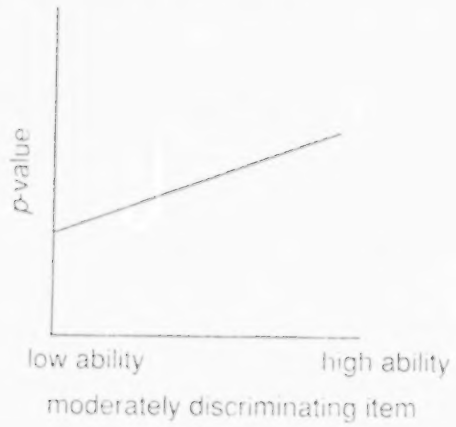

Figure 4. Characteristics of a moderately discriminating test item (Osterlind, 1989, p. 285).

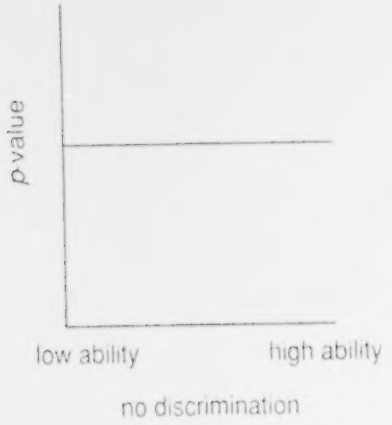

Figure 5. Characteristics of a nondiscriminating test item (Osterlind, 1989, p. 286).

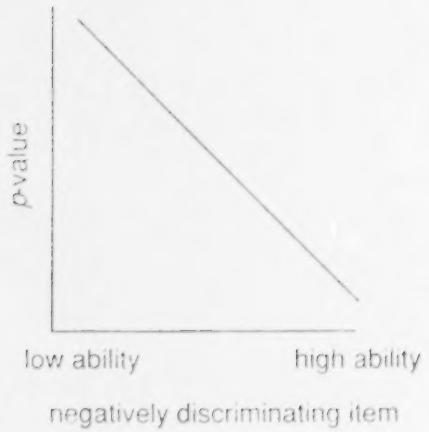

Figure 6. Characteristics of a negatively discriminating test item (Osterlind, 1989, p. 286).

where Uc is the mean criterion score for those examinees who answer the item correctly

Ux is the mean score for the entire group

σx is the standard deviation for the entire group

p is the item difficulty

q is (1-p).

The second problem is that the range of the statistic is dependent upon the difficulty of the item. One such example of a restricted range is observed when an item is either very easy or very difficult and there is very little if any differentiation between the high- and low-ability groups.

The point-biserial correlation is also used to assess the effectiveness of the incorrect options. For these distractors or foils, the point-biserial should be negative, suggesting that these options attract a greater number of low-performing students than high-performing students in question.

Biserial correlation. The biserial correlation coefficient is given by

$$\rho bis = \frac{Uc - Ux}{\sigma x} (p/Y)$$

where Uc is the criterion score mean of those examinees who answered the item correctly

Ux is the total test score mean of all the examinees

σx is the standard deviation of the total test score of all the examinees

p is the proportion of examinees who answered the item correctly

Y is the Y ordinate of the standard normal curve at the Z-score associated with the p-value of the item.

Mathematically, the relationship between the biserial and point-biserial correlation can be depicted as

$$\rho bis = \frac{\sqrt{pq}}{Y} \rho pbis$$

"Because the value of the Y ordinate on a normal curve is always less than $\sqrt{pq}$, the value of a biserial correlation will always be at least one-fifth greater than the point-biserial correlation for the same variables" (Crocker & Algina, 1986, p. 318). This difference remains fairly moderate for items of average difficulty, but as p-values drop below .25 and rise above .75, the difference between biserial and point-biserial correlations increases sharply, anything up to as much as four times greater for biserial compared to point-biserial correlation.

In item analysis, the main advantage of the biserial correlation lies in its ability to overcome the dependency of the point-biserial correlation upon the difficulty level. The biserial correlation range may exceed 1, but this hardly occurs when it can be assumed that the construct being assessed is normally distributed among examinees.

However, the choice of which of these two indices to use should be made on the basis of the nature of the response. If it is a true dichotomy, then the point-biserial correlation is used; but if it is a forced or artificial dichotomy, then the biserial correlation is used.

The computations of these two indices were completed using LERTAP.

Estimating item discriminating power (D). A simple but practical way of estimating the item discriminating power (D) is to subtract the number of examinees in the low-ability group who answered the item correctly from the number in the upper-ability group who answered the item correctly and then divide by the number in each group:

$$D = \frac{RU - RL}{\frac{1}{2}T}$$

where D is the index of the item discriminating power

RU is the number in the upper group who answered the item correctly

RL is the number in the lower group who answered the item correctly

½T is one half of the total number of examinees included in the item analysis group.

Several rules exist for forming the two groups. The most common are the top and bottom 27% or the top and bottom half. As for the point-biserial and biserial correlation, the D index may be used to assess the effectiveness of the distractors. In these cases the D value should be negative.

Item Analysis Using Item Response Theory

The construction and development of tests could be considerably improved by using additional information from item responses. In particular, classical test score model item analysis statistics do not provide information about how examinees at different ability levels on the trait have performed on the item. Item response theory (IRT), which consists of a family of models such as the one-, two-, and three-parameter models, can be of immense help in the design, construction, and evaluation of educational tests.

For this study only the one-parameter Rasch model was used to analyse data. Below is a description of the main features of item response theory in general.

Item Parameters in Item Response Theory (IRT)

Using item response theory (IRT), it is possible to graph item trace lines which depict information about one, two, or three parameters, or mathematical boundaries, for each item. According to Hambleton, Swaminathan, and Rogers (1991) and Osterlind (1989), such item trace lines are typically called item characteristic curves (ICCs). Typically, the parameters depicted are (a) Parameter A, indicating the "steepness" of the item trace line and representing the probability of responding correctly to an item, increasing as one goes up the scale as a measure of discrimination among varying ability groups; (b) Parameter B, defining the difficulty

of the item by noting the point on the ability scale at the point of inflection for the curve; (c) Parameter C, showing the beginning, or base, of the curve, suggesting the probability of guessing (also called "chance" or "pseudo-chance") a correct response on the item for very low-ability examinees.

The ICC plots "percent success" along the ordinate (Y axis) and the examinee attribute/ability along the abscissa (X axis). It is also important to observe that the slope of any curve is monotonic (that is, it always rises and is never exactly horizontal); that an "inflection point" (which can be shown by drawing a horizontal line from a point on the curve to the Y axis) is determined by the left-to-right shift of the curve; and that the two asymptotes, lower and upper, may approach but never actually reach 0.00 (for the one- and two-parameter models; C for the three-parameter model) and 1.00, respectively. An ICC is therefore technically termed a *monotonic normal ogive*. Ogives are merely specialised graphical representations of a frequency distribution.

According to Osterlind (1989), discrimination, difficulty, and guessing for an item can simultaneously be displayed graphically, thus making ICCs especially powerful devices for analysing items. An examination of Figures 7 to 9 illustrates some of the major item differences.

Figure 7 displays an ICC curve for a single item. Such an ICC serves to provide a model for a typical item. Lord (1952) developed the equation for the ICC in terms of populations using normal curve theory. However, the mathematics to get estimates for the parameters in the model became intractable. Subsequently, it was recognised by Birnbaum (cited in Lord & Novick, 1968) that the logistic curve was a very good model to the normal curve. The mathematics associated with the logistic curve could be performed. The equation below is for an ICC displayed in Figure 7.

$$P(ujg/\theta) = Cg + (1-Cg)\frac{eDag\,(\theta-bg)}{1+eDag\,(\theta-bg)}$$

where $P(ujg/\theta)$ is the probability that a randomly selected person j with ability θ will respond correctly to an item g;

ag is the item discrimination;

bg is the item difficulty;

Cg is the level of guessing

D is the scaling factor, generally set to 1.7, and e is the base of the natural logarithm.

Now, if there is no guessing, $Cg=0$. Further, if it can be assumed that item discrimination is the same for all items, then,

$$a1 = a2 = \ldots = ak = \tilde{a}.$$

Accepting these two observations yields the one parameter Rasch model employed in the present study. Under these assumptions, the equation becomes:

$$Pg(\theta) = \frac{eD\,(\theta-bg)}{1+eD\,(\theta-bg)}$$

where $Pg(\theta)$ is the probability that a randomly chosen examinee with ability θ answers item g correctly;

bg is the item difficulty;

D is the sealing factor, generally set to 1.7, and e is the base of the natural logarithm.

Figure 8 displays ICCs for two items which are similar in many respects but differ in difficulty. Items 1 and 2 have similar shapes, indicating that the two items discriminate at about the same rate; however, because the curve for item 2 is shifted further to the right than that for item 1, item 2 discriminates at a higher level of ability than does item 1. Item 2 can be described as being more difficult in

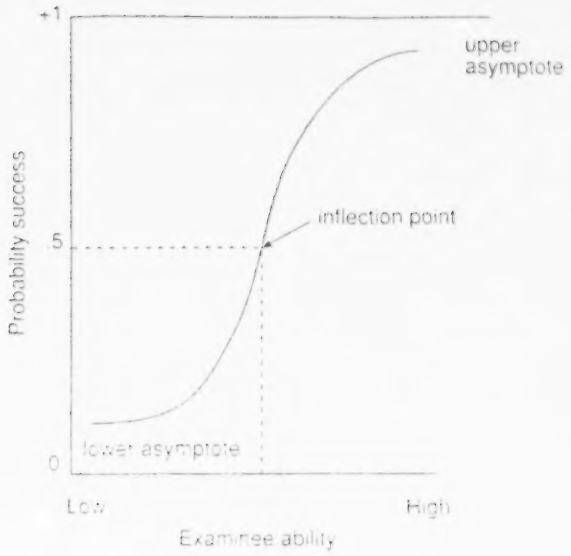

Figure 7. Trace line for a single item (Osterlind, 1989, p. 296).

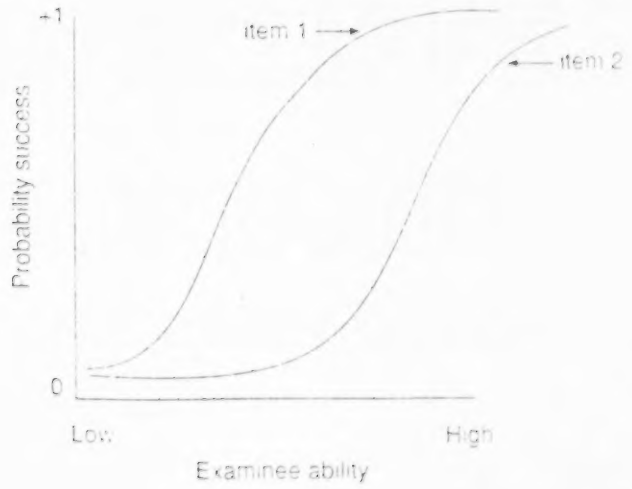

Figure 8. ICC of two similar items of different difficulty (Osterlind, 1989, p. 297).

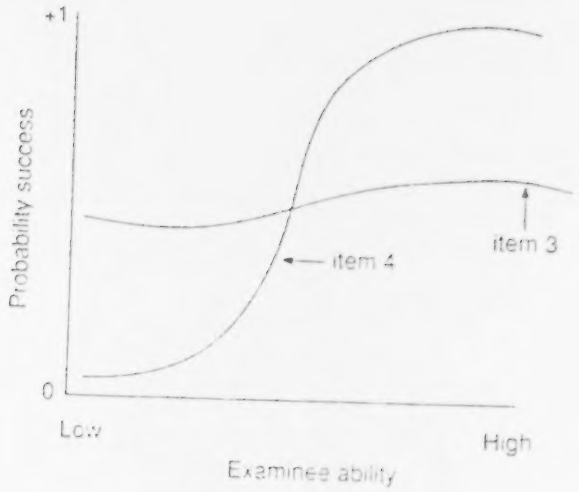

Figure 9. ICCs for two different items (Osterlind, 1989, p. 298).

comparison to item 1. Such item qualities can be gleaned just by examining the ICC and making sure that appropriate items are selected for appropriate ability levels/groups.

Figure 9 shows ICCs for items 3 and 4. The ICC for item 3 is very flat, indicating that the item discriminates equally across the abilities of all the examinees. Items with similar ICCs do not serve a useful purpose if the intention is to show differences across ability levels. Item 4 has a very steep slope, but at only one point (the average ability level) along the ability continuum. Such ICCs are appropriate under circumstances where sharp ability differentials are required among examinees of average ability, but test developers tend to favour items whose ICC is of the smooth, lazy-S form shown in Figure 8.

Issues of Validity

Content and cognitive representativeness. As pointed out earlier, the 1987 ZJC Geography Specimen Paper 1 was not constructed according to a full table of specifications. Consequently, one of the tasks undertaken in the present study was to assess the representativeness that did result in terms of a content-by-cognitive level specification table. Such a table was developed by the researcher at a later stage for use during the construction of actual examination papers at the ZJC level. The content dimension included four topics: physical geography, economic geography, and population and settlement, in addition to mapwork. The cognitive dimension included three levels: knowledge with understanding, comprehension of information and ideas, and the recall and application of skills. A copy of the specifications table is included in Appendix C.

A panel of 15 judges was used to classify the 50 items into the four content areas and three cognitive levels shown in the table of specifications. All 15 judges were familiar with the ZJC syllabus and had taught or were still teaching at this level. All had had experience in constructing geography multiple-choice items either at the ZJC or the ordinary level. Thirteen of the judges were university geography graduates; the remaining two had geography diplomas from teacher-training colleges.

Before the judges were left to work on their own, they underwent a half-day training and practice session on the classification of test items according to the two dimensions. The extent to which the judges agreed on the content areas represented the degree to which the items could be considered as sampling the four subject domains, and the extent to which they agreed on the cognitive levels represented the degree to which the items could be regarded as assessing the three cognitive levels.

The level of difficulty of each subtest (including a single item) was assessed in an attempt to probe what a typical test in geography at the ZJC level is like. An additional but important desire which the researcher had in respect of test-item

difficulty was the extent to which mapwork items affected individuals' scores, especially in view of the opinion held in Zimbabwe that mapwork is a difficult skill for many students.

Test-wiseness. The concept of test-wiseness and its effect on the validity of test scores has been the subject of interest in many studies (Benson, 1988; Messick, 1989; Millman, 1966; Rogers & Bateson, 1991; Rogers & Wilson, 1993). In fact, it should not take anybody some special understanding to notice that when certain test items contain obvious clues to the answers, a fundamental validity problem exists. The researcher was therefore particularly interested in probing the extent to which the notion of test-wise susceptibility was prevalent in educational tests in Zimbabwe, especially in view of the important decisions that were often made on the basis of individual test scores. In attempting to probe the effect of test-wiseness on scores for individual students, attention was focussed on those candidates who could be described as high-ability and low-ability groups. The assumption made was that the high-ability group would be able to exploit the situation arising out of any test-wise item, whereas the low-ability group would be less able to do so.

The strategies used to identify items susceptible to test-wiseness were taken from Millman et al. (1965) and represent the most common test-wise cues found in school-leaving multiple-choice examinations.

1. Four deductive-reasoning strategies:

   ID1 - Eliminate options known to be incorrect.

   ID2 - Choose neither or both of two options which imply the correctness of each other.

   ID3 - Choose neither or one of two options, one of which, if correct, would imply the incorrectness of the other.

   ID5 - Utilise relevant content information in other test items and options.

2. One cue-using strategy:

IIB4 - Recognise and use similarities between the stem and the options.

The panel of 15 judges who assessed the content-by-cognitive representativeness of the ZJC specimen paper was asked to indicate separately which items they considered test-wise susceptible. A copy of the instructions provided to the judges is provided in Appendix C. The consensus on test-wise items was also subjected to two independent opinions, those of a subject expert and a psychometrician. In addition, consensus among the judges was defined as 54% or more agreement that an item contained an identified test-wise element.

The judgments of the panel were then examined in terms of the actual examinee-response pattern for each item option as determined in the LERTAP analysis. Items were classified as test-wise susceptible if the p-value for an option attracted less than 5% of the examinees.

Gender differences. The last major research question revolved around gender differences in performance. In particular, was there any noticeable difference between the performance of boys and that of girls? The difference between average or mean scores and the difference between item p-values for boys and girls were examined to address this question.

Chapter 4

Presentation and Analysis of Results

The results of the study are presented in this chapter in three parts. The first part contains a summary of the students' performance. The correlations among the subtests, the total test, and gender are presented in the second part. Issues of validity are addressed in the third and final part.

In attempting to address the issue of validity, a number of approaches were adopted. In the first place, empirical evidence from students' responses in the form of item difficulty (p-values) and item discrimination (point-biserial correlation) was combined with interrater agreement in order to probe the notion of test-wiseness. In the second place, judges' agreement was used to validate the item content. Lastly, gender differences were probed by examining the item and total-score means for boys, on one hand, and for girls, on the other. A t-test was also carried out to determine the significance of the difference, if any, between the mean score for boys and that for girls.

Description of Students' Performance

Presented in Table 7 are the number of items, mean (x), standard deviation (SD), internal consistency coefficient (r), and standard error of measurement (SEM) for each of the four subtests and the total test. As well, Cronbach's alpha for the composite formed by aggregating the four subtest scores reported is provided. To facilitate comparisons in performance among the subtests and the total test, the summary values are reported as percentages, as well as in the raw score metric.

Figures 10-14 are histograms showing the distribution of scores for the four subtests and the total test. For the total test (Figure 14) the scores show a near-normal distribution, with a mean of 52.9 and a standard deviation of 6.72. The highest and lowest scores are 47 and 8, respectively, out of 50. Inspection of

Table 7

Summary of Student Performance on Four Subtests and Total Test

Subtest	# of items	HS	LS	x	%	SD	%	(r)	SEM	%
Mapwork	12	10.00	0.00	4.54	37.8	2.01	16.8	0.39	1.51	12.6
Physical geography	14	14.00	1.00	7.47	53.4	2.41	17.2	0.55	1.56	11.1
Economic geography	19	19.00	3.00	11.50	60.5	3.00	15.8	0.61	1.83	9.6
Population and settlement	5	5.00	0.00	2.95	59.0	1.26	25.2	0.46	0.83	16.6
Total test (Hoyt)	50	47.00	8.00	26.43	52.0	6.72	13.4	0.79	3.06	6.1
Total test Cronbach's alpha for a composite	4 subtests							0.73[a]		

HS = Highest score LS = Lowest score

[a] Cronbach's composite alpha

Z	LB	UB	F	P	CF	CP	EACH * REPRESENTS 3 OBSERVATIONS
-8.21	-14.50	-13.50	0	0.00	0	0.00	
-8.71	-13.50	-12.50	0	0.00	0	0.00	
-8.22	-12.50	-11.50	0	0.00	0	0.00	
-7.72	-11.50	-10.50	0	0.00	0	0.00	
-7.22	-10.50	-9.50	0	0.00	0	0.00	
-6.73	-9.50	-8.50	0	0.00	0	0.00	
-6.23	-8.50	-7.50	0	0.00	0	0.00	
-5.73	-7.50	-6.50	0	0.00	0	0.00	
-5.24	-6.50	-5.50	0	0.00	0	0.00	
-4.74	-5.50	-4.50	0	0.00	0	0.00	
-4.24	-4.50	-3.50	0	0.00	0	0.00	
-3.75	-3.50	-2.50	0	0.00	0	0.00	
-3.25	-2.50	-1.50	0	0.00	0	0.00	
-2.75	-1.50	-0.50	0	0.00	0	0.00	
-2.26	-0.50	0.50	8	0.88	8	0.88	***
-1.76	0.50	1.50	37	4.07	45	4.95	************
-1.26	1.50	2.50	98	10.78	143	15.73	**********************************
-0.76	2.50	3.50	163	17.93	306	33.66	**
-0.27	3.50	4.50	168	18.48	474	52.15	**
0.23	4.50	5.50	147	16.17	621	68.32	***
0.73	5.50	6.50	123	13.53	744	81.85	***
1.22	6.50	7.50	96	10.56	840	92.41	**********************************
1.72	7.50	8.50	41	4.51	881	96.92	**************
2.22	8.50	9.50	22	2.42	903	99.34	*******
2.71	9.50	10.50	6	0.66	909	100.00	**
3.21	10.50	11.50	0	0.00	909	100.00	
3.71	11.50	12.50	0	0.00	909	100.00	
4.20	12.50	13.50	0	0.00	909	100.00	
4.70	13.50	14.50	0	0.00	909	100.00	
5.20	14.50	15.50	0	0.00	909	100.00	
5.70	15.50	16.50	0	0.00	909	100.00	
6.19	16.50	17.50	0	0.00	909	100.00	
6.69	17.50	18.50	0	0.00	909	100.00	
7.19	18.50	19.50	0	0.00	909	100.00	
7.68	19.50	20.50	0	0.00	909	100.00	
8.18	20.50	21.50	0	0.00	909	100.00	
8.68	21.50	22.50	0	0.00	909	100.00	
9.17	22.50	23.50	0	0.00	909	100.00	
9.67	23.50	24.50	0	0.00	909	100.00	
10.17	24.50	25.50	0	0.00	909	100.00	

Figure 10. Mapwork scores.

Z	LB	UB	F	P	CF	CP	EACH * REPRESENTS 3 OBSERVATIONS
-8.07	-12.50	-11.50	0	0.00	0	0.00	
-7.65	-11.50	-10.50	0	0.00	0	0.00	
-7.24	-10.50	-9.50	0	0.00	0	0.00	
-6.82	-9.50	-8.50	0	0.00	0	0.00	
-6.41	-8.50	-7.50	0	0.00	0	0.00	
-6.00	-7.50	-6.50	0	0.00	0	0.00	
-5.58	-6.50	-5.50	0	0.00	0	0.00	
-5.17	-5.50	-4.50	0	0.00	0	0.00	
-4.75	-4.50	-3.50	0	0.00	0	0.00	
-4.34	-3.50	-2.50	0	0.00	0	0.00	
-3.92	-2.50	-1.50	0	0.00	0	0.00	
-3.51	-1.50	-0.50	0	0.00	0	0.00	
-3.10	-0.50	0.50	0	0.00	0	0.00	
-2.68	0.50	1.50	5	0.55	5	0.55	**
-2.27	1.50	2.50	16	1.76	21	2.31	*****
-1.85	2.50	3.50	38	4.18	59	6.49	************
-1.44	3.50	4.50	46	5.06	105	11.55	**************
-1.02	4.50	5.50	72	7.92	177	19.47	************************
-0.61	5.50	6.50	134	14.74	311	34.21	**
-0.20	6.50	7.50	145	15.95	456	50.18	**
0.22	7.50	8.50	131	14.41	587	64.56	***
0.63	8.50	9.50	122	13.42	709	78.00	**
1.05	9.50	10.50	108	11.88	817	89.88	**********************************
1.46	10.50	11.50	63	6.93	880	96.81	********************
1.87	11.50	12.50	19	2.09	899	98.90	******
2.29	12.50	13.50	7	0.77	906	99.67	**
2.70	13.50	14.50	3	0.33	909	100.00	*
3.12	14.50	15.50	0	0.00	909	100.00	
3.53	15.50	16.50	0	0.00	909	100.00	
3.95	16.50	17.50	0	0.00	909	100.00	
4.36	17.50	18.50	0	0.00	909	100.00	
4.77	18.50	19.50	0	0.00	909	100.00	
5.19	19.50	20.50	0	0.00	909	100.00	
5.60	20.50	21.50	0	0.00	909	100.00	
6.02	21.50	22.50	0	0.00	909	100.00	
6.43	22.50	23.50	0	0.00	909	100.00	
6.84	23.50	24.50	0	0.00	909	100.00	
7.26	24.50	25.50	0	0.00	909	100.00	
7.67	25.50	26.50	0	0.00	909	100.00	
8.09	26.50	27.50	0	0.00	909	100.00	

Figure 11. Physical geography scores.

Z	LB	UB	F	P	CF	CP	EACH * REPRESENTS 2 OBSERVATIONS
-6.50	-8.50	-7.50	0	0.00	0	0.00	
-6.17	-7.50	-6.50	0	0.00	0	0.00	
-5.83	-6.50	-5.50	0	0.00	0	0.00	
-5.50	-5.50	-4.50	0	0.00	0	0.00	
-5.17	-4.50	-3.50	0	0.00	0	0.00	
-4.83	-3.50	-2.50	0	0.00	0	0.00	
-4.50	-2.50	-1.50	0	0.00	0	0.00	
-4.17	-1.50	-0.50	0	0.00	0	0.00	
-3.83	-0.50	0.50	0	0.00	0	0.00	
-3.50	0.50	1.50	0	0.00	0	0.00	
-3.17	1.50	2.50	0	0.00	0	0.00	
-2.83	2.50	3.50	6	0.66	6	0.66	***
-2.50	3.50	4.50	8	0.88	14	1.54	****
-2.17	4.50	5.50	19	2.09	33	3.63	**********
-1.83	5.50	6.50	20	2.20	53	5.83	**********
-1.50	6.50	7.50	41	4.51	94	10.34	**********
-1.17	7.50	8.50	53	5.83	147	16.17	**********
-0.83	8.50	9.50	78	8.58	225	24.75	**********
-0.50	9.50	10.50	100	11.00	325	35.75	**********
-0.17	10.50	11.50	101	11.11	426	46.86	**********
0.17	11.50	12.50	113	12.43	539	59.30	**********
0.50	12.50	13.50	126	13.86	665	73.16	**********
***							
0.83	13.50	14.50	101	11.11	766	84.27	**********
1.17	14.50	15.50	71	7.81	837	92.08	**********
1.50	15.50	16.50	46	5.06	883	97.14	**********
1.83	16.50	17.50	15	1.76	899	98.90	********
2.17	17.50	18.50	8	0.88	908	99.85	*****
2.50	18.50	19.50	1	0.11	909	100.00	*
2.83	19.50	20.50	0	0.00	909	100.00	
3.17	20.50	21.50	0	0.00	909	100.00	
3.50	21.50	22.50	0	0.00	909	100.00	
3.83	22.50	23.50	0	0.00	909	100.00	
4.17	23.50	24.50	0	0.00	909	100.00	
4.50	24.50	25.50	0	0.00	909	100.00	
4.83	25.50	26.50	0	0.00	909	100.00	
5.17	26.50	27.50	0	0.00	909	100.00	
5.50	27.50	28.50	0	0.00	909	100.00	
5.83	28.50	29.50	0	0.00	909	100.00	
6.17	29.50	30.50	0	0.00	909	100.00	
6.50	30.50	31.50	0	0.00	909	100.00	

Figure 12. Economic geography scores.

Z	LB	UB	F	P	CF	CP	EACH * REPRESENTS 4 OBSERVATIONS
-15.88	-17.50	-16.50	0	0.00	0	0.00	
-15.08	-16.50	-15.50	0	0.00	0	0.00	
-14.28	-15.50	-14.50	0	0.00	0	0.00	
-13.49	-14.50	-13.50	0	0.00	0	0.00	
-12.69	-13.50	-12.50	0	0.00	0	0.00	
-11.90	-12.50	-11.50	0	0.00	0	0.00	
-11.10	-11.50	-10.50	0	0.00	0	0.00	
-10.31	-10.50	-9.50	0	0.00	0	0.00	
-9.51	-9.50	-8.50	0	0.00	0	0.00	
-8.71	-8.50	-7.50	0	0.00	0	0.00	
-7.92	-7.50	-6.50	0	0.00	0	0.00	
-7.12	-6.50	-5.50	0	0.00	0	0.00	
-6.33	-5.50	-4.50	0	0.00	0	0.00	
-5.53	-4.50	-3.50	0	0.00	0	0.00	
-4.73	-3.50	-2.50	0	0.00	0	0.00	
-3.94	-2.50	-1.50	0	0.00	0	0.00	
-3.14	-1.50	-0.50	0	0.00	0	0.00	
-2.35	-0.50	0.50	34	3.74	34	3.74	*********
-1.55	0.50	1.50	96	10.56	130	14.30	**********
-0.76	1.50	2.50	183	18.58	209	32.89	**********
0.04	2.50	3.50	273	30.03	572	62.93	**********
********							
0.84	3.50	4.50	257	28.27	829	91.20	**********
****							
1.63	4.50	5.50	80	8.80	909	100.00	**********
2.43	5.50	6.50	0	0.00	909	100.00	
3.22	6.50	7.50	0	0.00	909	100.00	
4.02	7.50	8.50	0	0.00	909	100.00	
4.82	8.50	9.50	0	0.00	909	100.00	
5.61	9.50	10.50	0	0.00	909	100.00	
6.41	10.50	11.50	0	0.00	909	100.00	
7.20	11.50	12.50	0	0.00	909	100.00	
8.00	12.50	13.50	0	0.00	909	100.00	
8.79	13.50	14.50	0	0.00	909	100.00	
9.59	14.50	15.50	0	0.00	909	100.00	
10.39	15.50	16.50	0	0.00	909	100.00	
11.18	16.50	17.50	0	0.00	909	100.00	
11.98	17.50	18.50	0	0.00	909	100.00	
12.77	18.50	19.50	0	0.00	909	100.00	
13.57	19.50	20.50	0	0.00	909	100.00	
14.37	20.50	21.50	0	0.00	909	100.00	
15.16	21.50	22.50	0	0.00	909	100.00	

Figure 13. Population and settlement scores.

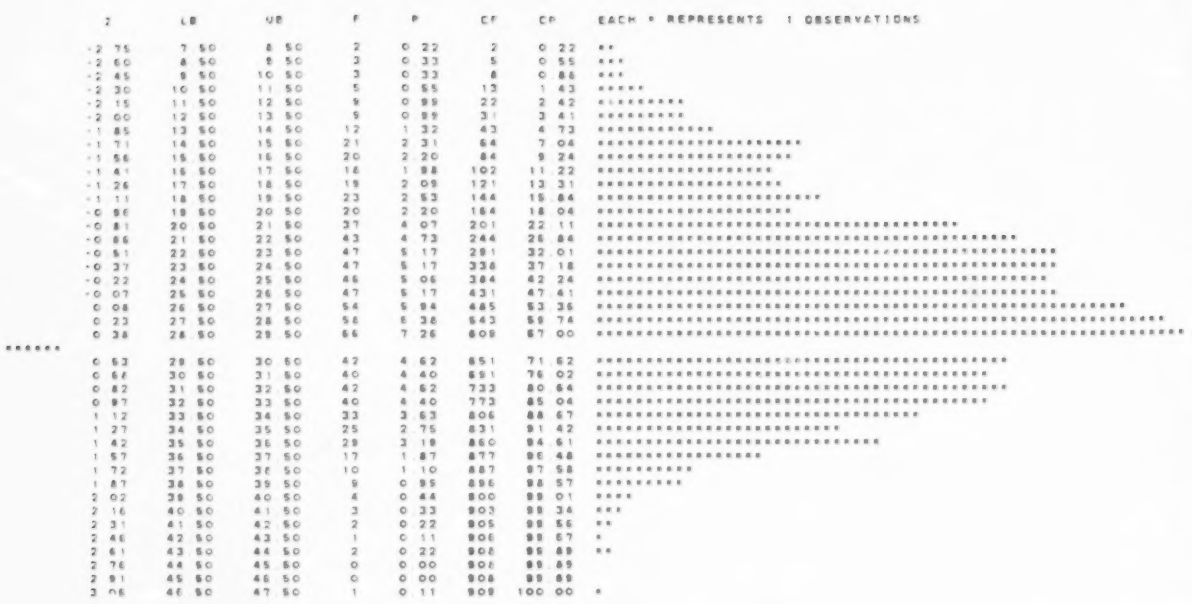

Figure 14. Total test scores.

Figure 10 reveals that mapwork scores were essentially positively skewed, with a mean of 37.8%, a standard deviation of 16.8%, and with the highest and lowest scores at 83.3% and 0.00%, respectively, out of 12, indicating a difficult subtest. The distribution of scores for the physical geography subtest is almost normal, with a mean of 53.4%, a standard deviation of 17.2%, and the highest score of 100% and lowest score of 71% out of 14. Economic geography, which is the easiest subtest, has a slightly negatively skewed distribution, with a mean of 60.5%, a standard deviation of 15.8%, and the highest and lowest scores at 100% and 15.8%, respectively. Population and settlement showed a slightly positively skewed distribution, with a mean of 59.0%, a standard deviation of 25.2%, the highest score at 100%, and the lowest at 0%. Taken together, the results for the four subtests indicate that the subtests were of unequal difficulty, with economic geography and population and settlement being the easiest and mapwork the most difficult.

Description of Test Characteristics

Internal consistency. The internal consistency estimates and the corresponding standard error of measurement are reported in Table 7. The total test reliability of 0.79 (see Table 7) reveals a reasonably high reliability index to suggest the test taken as a whole has an acceptable level of internal consistency. Most measurement specialists (Ary et al., 1990; Crocker & Algina, 1986; Ebel & Frisbie, 1991; Gronlund, 1993; Hopkins & Stanley, 1990; Mehrens & Lehmann, 1991; Sax, 1989) quoted reliabilities of .70 and above as indicating a test with a sound cohesion. The standard error of measurement is 6.1% or approximately 3 points out of a total of 50.

Table 7 also shows that Cronbach's alpha for the composite, 0.73, was lower than the internal consistency for the total test, 0.79. The difference is attributable to the differences in the unit of analysis employed in getting the two estimates. The total test reliability was estimated using Hoyt's (1941) ANOVA procedures, in which the analysis is conducted at the item level. In contrast, Cronbach's alpha for a composite is based on an analysis at the subtest level (Feldt & Brennan, 1989).

Consistent with existing literature and research findings (Ary et al., 1990; Crocker & Algina, 1986; Cronbach, 1990; Ebel & Frisbie, 1991; Feldt, 1993; Feldt & Brennan, 1989; Grier, 1975; Smith & Glass, 1987), internal consistency of the subtests increased as the number of test items increased, with one exception. The subtest on population and settlement, with only five items, technically should have reflected the lowest coefficient. However, this subtest had a reliability coefficient of 0.46, whereas mapwork, with 12 items, had the lowest reliability coefficient, 0.39. Such a finding is not at all surprising given the argument that the reliability of a test is in part a function of many factors, such as examinee ability, the variability of the sample of examinees, and the nature of the construct being measured (Ary et al, 1990; Smith & Glass, 1987). In the case at hand, as indicated above, the mapwork subtest was the most difficult test. The subtest involved an assessment in which the

students were asked to interpret an actual map, whereas the population and settlement subtest involved assessment of the students' abilities to use abstractions of reality or models.

Relationship among subtests. The intercorrelations among the subtests, total test, and gender are reported in Table 8. There were moderate relationships between the observed scores for (a) mapwork and physical geography (0.42); (b) mapwork and economic geography (0.41); (c) mapwork and population and settlement (0.35); (d) physical geography and population and settlement (0.44); (e) economic geography and population and settlement (0.49); and (f) physical and economic geography (0.52).

Table 8

Correlations Between Subtest, Total Test, and Gender

	1	2	3	4	5	6
Variable	MPWK	PHYGEOG	ECONGEOG	POPSET	TT	EI
1. MPWK	1.00					
2. PHYGEOG	0.42	1.00				
3. ECONGEOG	0.41	0.52	1.00			
4. POPSET	0.35	0.44	0.49	1.00		
5. TT	0.70	0.80	0.85	0.67	1.00	
6. EI	0.08	0.14	0.21	0.09	0.18	1.00

1. MPWK = Subtest 1, mapwork
2. PHYGEOG = Subtest 2, physical geography
3. ECONGEOG = Subtest 3, economic geography
4. POPSET = Subtest 4, Population and settlement
5. TT = Total Test
6. EI = Gender

Item Characteristics

Item difficulty. Table 9 shows the difficulty level for each item. In keeping with the earlier finding that, at the subtest level, mapwork had the lowest mean, the difficulty indices for the mapwork items tend to be the lowest. Indeed, the most difficult item (item 5), where 84% of the examinees failed to get the correct answer, is in this subtest. Mapwork also had the lowest proportion of items (25.0%), which half or more of the examinees correctly answered. In fact, even on these three items where 50% or more got the correct answer, the proportion that did so ranged from just 50% to 56%. By contrast, economic geography, which was a relatively easier subtest, contained the easiest item (item 41), in which only 11% of all the examinees failed to get the correct answer. Further, at least 70% of the examinees correctly answered 14 of the 19 items in the economic geography subtest.

These results indicate how, by conducting an analysis at the item level in a well-conducted pilot study, it would be possible to construct a final test at a desired level of difficulty (for example, 60% uniform across topic [subtest]).

Figures 15, 16, and 17 show plots for the easiest, moderately difficult, and most difficult items on the test according to the levels of five ability groups from the lowest to the highest scores for a cohort of 20% for each group. For the most difficult item (item 5), all the examinees performed at or below chance (p-value .20 or below); for the moderately difficult item (item 34), 46% of the low-ability and 64% of the high-ability groups correctly answered the item; whilst for the easiest item (item 41), even 70% of the lowest ability group obtained the correct answer.

Item discrimination. Table 9 also shows point-biserial correlations (item-discriminating power). When recommendations made by Hopkins et al. (1990) are applied to the ZJC geography test, 5 groups of items can be identified as follows:

1. 2 items with negative discrimination and considered unproductive for test reliability purposes (items 5 and 36);

Table 9

Item Difficulty Levels

Mapwork			Physical geography			Economic geography			Population and settlement		
Item	p-value	rpbis	Item	p-value	rpbis	Item	p-value	rpbis	Item	p-value	rpbis
1.	.469	0.20	13.	.728*	0.40	16.	.738*	0.36	37.	.816*	0.47
2.	.501*	0.41	14.	.495	0.28	18.	.796*	0.48	38.	.682*	0.39
3.	.198†	0.21	15.	.783*	0.39	19.	.778*	0.34	39.	.754*	0.47
4.	.418	0.37	17.	.200†	0.23	20.	.723*	0.24	40.	.303	0.18
5.	.156†	-0.04	28.	.545*	0.42	21.	.470	0.21	48.	.395	0.40
6.	.242	0.28	29.	.429	0.16	22.	.689*	0.18			
7.	.488	0.32	30.	.878*	0.32	23.	.762*	0.36			
8.	.325	0.29	31.	.421	0.30	24.	.695*	0.43			
9.	.559*	0.34	32.	.705*	0.42	25.	.272	0.38			
10.	.257	0.15	34.	.578*	0.13	26.	.788*	0.40			
11.	.517*	0.24	35.	.212†	0.17	27.	.507*	0.39			
12.	.408	0.35	43.	.208	0.32	33.	.694*	0.35			
			45.	.738*	0.39	36.	.263	-0.06			
			49.	.552*	0.40	41.	.891*	0.31			
						42.	.718*	0.27			
						44.	.535*	0.19			
						46.	.505*	0.32			
						47.	.495	0.37			
						50.	.178†	0.22			

* Indicates an item where 50% or more of the examinees got the correct answer.
† Indicates p-value at about or less than chance (.20 for 5-option item).

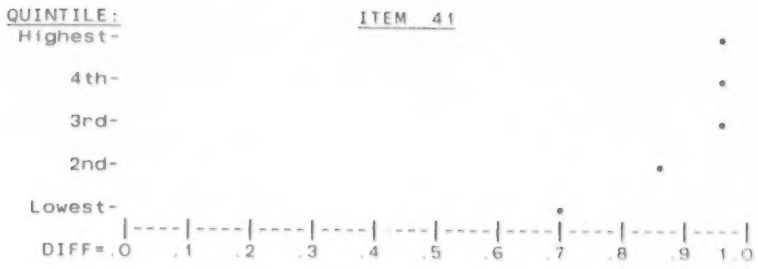

Figure 15. Plot for the easiest item.

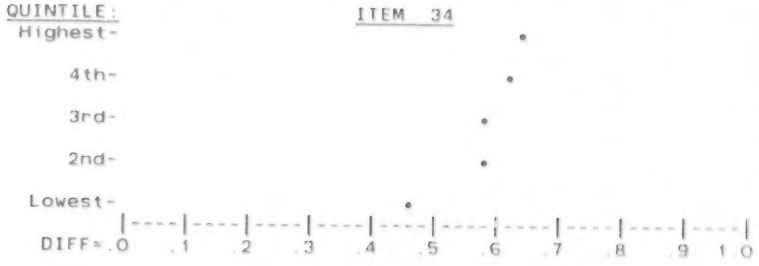

Figure 16. Plot for a moderately difficult item.

Figure 17. Plot for the most difficult item.

2. 9 marginally discriminating items with discrimination falling between .10 and .19 (items 6, 8, 10, 22, 29, 34, 35, 40, and 44);

3. 9 reasonably discriminating items with discrimination falling between .20 and .29 (items 1, 3, 11, 14, 17, 20, 21, 42, and 50);

4. 19 discriminating items with discrimination falling between .30 and .39 (items 4, 7, 9, 12, 15, 16, 19, 23, 25, 27, 30, 31, 33, 38, 41, 43, 45, 46, and 47);

5. 11 highly discriminating items with discrimination lying at .40 and above (items 2, 13, 18, 24, 26, 28, 32, 37, 39, 48, and 49).

In general, it has been observed (Hopkins et al., 1990) that the discriminating power of an item is at a maximum when its difficulty level is .5, that is, when one half of the examinees are able to answer the item correctly. In addition, very easy and difficult items may not discriminate well. It should, however, be noted that a difficulty level of .5 does not guarantee that the item will discriminate well; neither will the fact that an item is either very easy or very difficult. The crucial test for an item to discriminate well depends on whether the high scorers on the total test agree with the keyed option (Hopkins et al., 1990). This latter reasoning would explain some of the anomalies observed in Table 9; namely, why some of the items with p-values approximating .5 (items 10, 34, and 44) are less discriminating, and why the easiest item, 41, has a lower discriminating power than, say, item 26. The same reasoning would also explain why items 5 and 36 have a negative discrimination.

Application of IRT to the ZJC Geography Test

Presented in this section are selected statistical analyses yielded by BICAL (one-parameter item response model) which are intended to complement the conventional item analyses already presented above. These are (a) the test characteristic curve, (b) standard errors of measurement, and (c) item fit statistics. As pointed out earlier, certain assumptions have to be made under the one-parameter item response model; in

particular, that there is no guessing and that item discrimination indices are equal across all levels of ability of the examinees.

Test characteristic curve. Figure 18 shows the test characteristic curve for the 50 items on the ZJC geography test. The curve is a graph of raw scores achieved by all examinees (on the vertical) against ability scores (along the horizontal). Under the one-parameter model, it is assumed that item difficulty is the only item characteristic influencing examinee performance. The distribution of ability in logistics is set at zero with a standard deviation of one and the probability of getting the correct answer at 50%. On this distribution the higher scores are located at the higher end of the ability scale. For the present study the highest score was 47, with only one examinee achieving that score; whereas the lowest score was 8, with two examinees obtaining this score.

It should also be noted that because the mean ability is set at 0, with a standard deviation of 1 and the probability of getting the item correct at 50%, the easier items will have negative indices, whereas the more difficult ones will have positive values. The difficulty indices for items 41, 34, and 5 were -2.11, -0.19, and 1.96, respectively.

Standard errors of measurement. Running down the left side of the test characteristic curve are shown standard errors of measurement. Unlike under the classical test score model, where there is only one standard error of measurement for all items, the one-parameter model shows that the standard error of measurement changes with changing ability. It is greatest at the highest/lowest scores and smallest around the mean. In the case of this study, the standard error of measurement for the highest score was 0.62, 0.41 for the lowest score, and 0.32 for the mean score.

The fact that the standard error of measurement is not the same across all levels of ability is one of the greatest advantages of IRT over the classical test score model. What this means is that there is greater precision in measurement around the mean

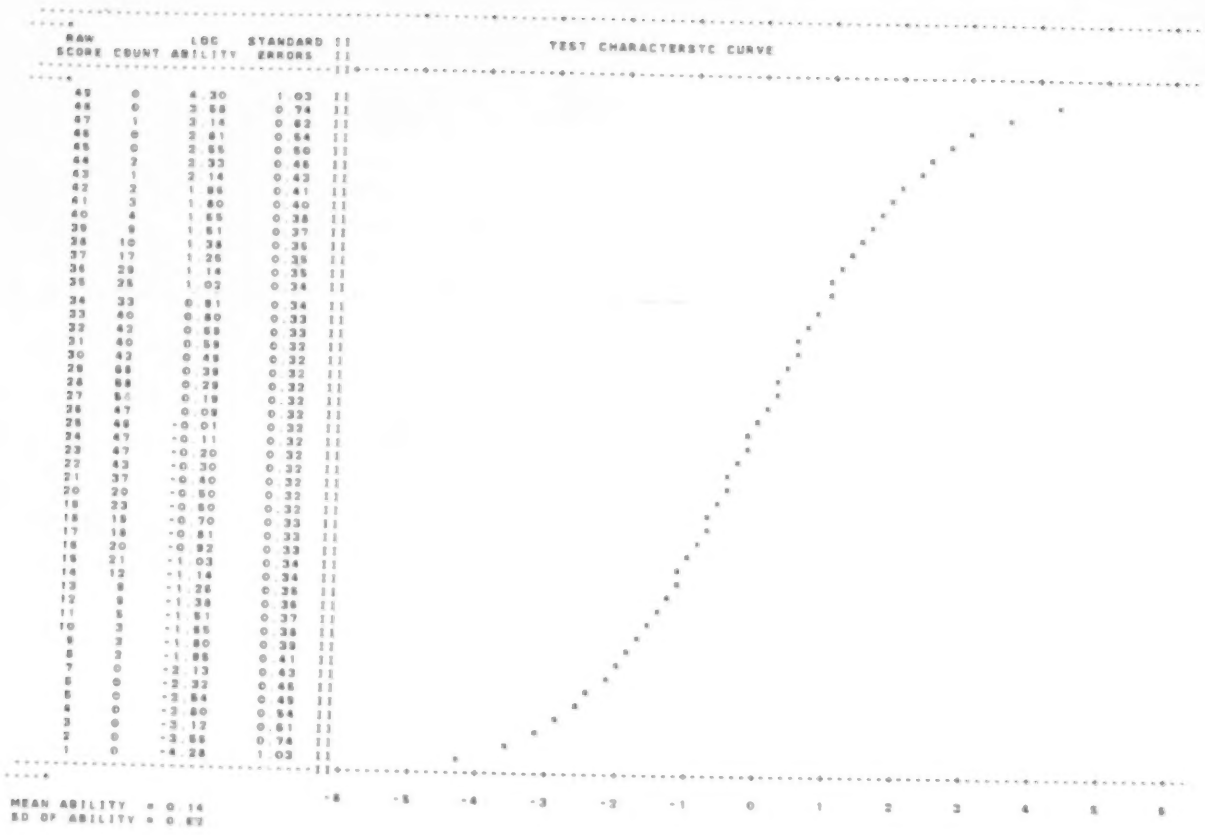

Figure 18. Test characteristic curve for the 50 items on the ZJC geography test.

than there is at the extreme scores. This should have implications where decisions such as selection have to be made on the basis of test scores.

Item fit statistics. Table 10 shows the item fit statistics for the easiest item (41), a moderately difficult item (34), and the most difficult item (5).

The item fit statistics are provided so that the adequacy of item difficulty across the total calibration sample (6 groups in this case) can be assessed. The groups are arranged according to the range of total raw scores obtained, with group 1 obtaining the lowest raw scores and group 6 obtaining the highest raw scores. BICAL uses item fit statistics based upon the analysis of residuals, namely, BETWEEN GROUP provided under "Item Characteristic Curve" and discrepancies between observed

Table 10

Item Fit Statistics for Items 5, 34, and 41

Item	Between group	Fit mean square	rpbis	Discrimination index
5	15.94	1.49	-0.04	-0.09
34	9.78	1.18	0.13	0.26
41	2.48	0.86	0.31	1.31

proportions under ICC and expected proportions. We would expect these discrepancies to be near one for a perfect fit, but because of sampling error, they are most likely to be different in reality. A rule of thumb is to eliminate items whose BETWEEN GROUP mean fit square is greater than 3.00 or 4.00 and whose point-biserial is below .25, or whose discrimination index lies outside 0.80 and 1.20. Applying these criteria to items in Table 10, we would need to revise items 5 and 34. Such a need is further confirmed when we examine ICCs for the three items shown in Figures 19-21. Figure 19 shows that item 5 is a very difficult item which is poorly discriminating across all levels of ability; Figure 20 shows that item 34 is a moderately difficult item which discriminates poorly at the higher level of ability; and while application of the statistical rules suggests that item 41 is functioning correctly, Figure 21 shows the item to be extremely easy and discriminating poorly, especially at the highest level of ability.

Description of Sources of Validity Evidence

Content-related evidence.

Item-topic classification. Table 11 provides a listing of the extent of agreement of item-topic classifications among the 12 members (out of the original 15 judges, responses were received from 12) of the panel of judges. As shown, there was almost unanimous agreement among the judges on the 12 items that made up the

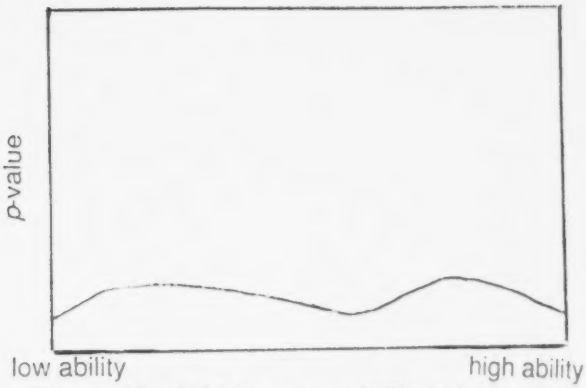

Figure 19. ICC for a very difficult item which also discriminates poorly (Item 5).

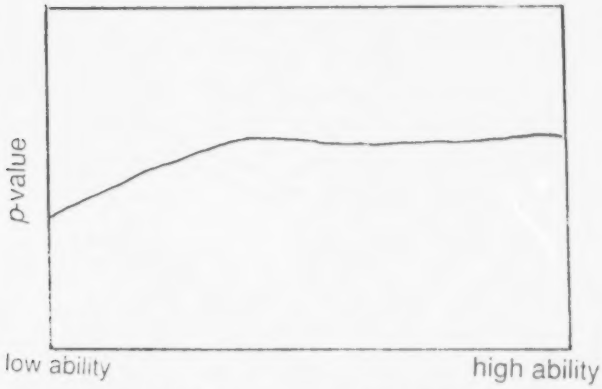

Figure 20. ICC for a moderately difficult item which discriminates poorly at higher ability (Item 34).

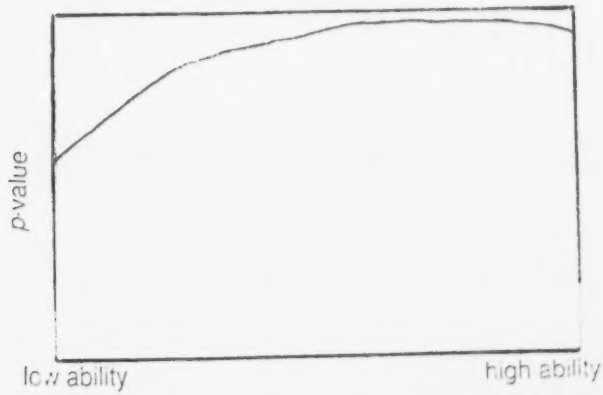

Figure 21. ICC for an easy item which discriminates poorly at the highest ability (Item 41).

Table 11

Item Content Classification by Judges

Question (item) classification by topic (content)			
Mapwork	Physical geography	Economic geography	Population & settlement
1:50000 map extract of headlands (24%)	Weather, climate, water cycle, and landform features (28%)	Water, energy, farming, manufacturing, and transport (38%)	(10%)
1 (100)	13 (100)	16 (83)	37* (75)
2 (100)	14 (92)	18* (58)	38 (83)
3 (100)	15 (100)	19 (83)	39* (75)
4 (100)	17* (75)	20 (92)	40 (92)
5 (100)	28 (83)	21 (92)	48 (83)
6 (100)	29 (83)	22 (92)	
7 (100)	30 (92)	23 (92)	
8 (100)	31 (92)	24 (92)	
9 (100)	32 (92)	25 (92)	
10 (92)	34* (67)	26 (92)	
11 (92)	35 (83)	27 (92)	
12 (100)	43 (83)	33 (92)	
	45* (67)	36 (83)	
	49 (83)	41 (92)	
		42* (58)	
		44 (83)	
		46 (92)	
		47 (83)	
		50* (67)	

1, 2 = Item (question) number
(58) (100) = % agreement between the 12 judges
* Denotes items where interrater agreement is less than 80%.

mapwork subtest (92%-100% agreement across items); and moderate to strong agreement on the 14 items in the physical geography subtest (67%-100%), the 19 items in the economic geography subtest (58%-92%), and the 5 items in the population and settlement subtest (75%-92%). It should be noted, however, that 80% agreement is the level most experts regard as the cut-off point for strong agreement. In the present case, 10 of the 12 judges would need to agree on a particular item-topic classification for strong agreement. Seven of the 50 items fail to meet this criterion: items 7, 34, 37, 39, 42, 45, and 50. Of these, three are within one judge of strong

agreement. In the case of the remaining two, items 34 and 45 are found in the physical geography section. That these two items showed a lower agreement resulted from the fact that they can be regarded as belonging to more than one distinct topic area. The remaining three items (18, 42, and 50) belong to the economic geography section. That these three items showed a lower agreement was not surprising because the concept of economic geography embraces a variety of activities which are not always discrete disciplines in and of themselves. It would be wise to revise these latter five items to ensure that they reflect one dominant topic area.

Content representativeness. In terms of the content representativeness, the distribution of items across the four topics was predetermined by the National Geography Panel which developed the test. From the subject expert's point of view, economic geography has the largest number of concepts to be studied and should therefore have had the highest number of items (38%). Physical geography is considered to be the base (foundation) upon which all geographical concepts are rooted and rightfully deserved second place (28%). Mapwork, which was third with 24% of the items, represents another important area where the geographer's unique language of communicating is exercised. Population and settlement, which had the smallest number of items (10%), achieved that position largely as a result of the limited number of different concepts on which to build test items.

Cognitive complexity. Attempts by the 12 panel members to classify the items according to cognitive levels as laid out in the *Taxonomy of Educational Objectives, Cognitive Domain* (Bloom et al., 1956) produced no meaningful pattern. Problems of attempting to classify test items according to cognitive abilities have been well documented. Wolf, Kelson, and Silver (1990) demonstrated how elusive the notion of "skills" can be when one tries to measure them in practical contexts. Of particular concern are difficulties that appear to arise due to the hierarchical and cumulative nature of Bloom's Taxonomy and a possible lack of agreement between what thinking

levels the item writers believe the students are using and what the students actually use. Gierl (1993), for example, found that the initial classification of test items into the levels of knowledge, comprehension, and application did not match the strategies used by Grade 6 students when responding to the test items.

The influence of test-wiseness. Presented in Appendix D is a detailed item analysis of each of the 50 items included in the ZJC geography test. The items are numbered 1-50 but are arranged according to the subtest to which they belong. The options (A-E) are shown with the keyed option asterisked (*). For each of the five options the item difficulty (p-value) and item discrimination (point-biserial correlation, rpbis) are shown. Where a test-wise item has been detected, the nature of the test-wiseness cue is indicated and a comment included to explain the pattern of the examinees' responses. The test-wise strategies used in the table are taken from Millman et al. (1965, p. 707) and include the following (see Appendix E):

ID1 absurd or nonfunctioning options

ID2 similar options

ID3 opposite or different options

ID5 utilise relevant content information in other test items and options

IIB4 stem-option link

Tables 12, 13, and 14 contain summaries of the test-wise-susceptible items and test-wise elements detected from the responses of the 909 students tested.

Of the 15 original judges, responses were received from 12, with responses from one of the judges being thrown out as representing an outlier in comparison to the responses from the other 11 judges. In addition, responses from a psychometrician and subject specialist were considered together with those from 11 judges, to give a total of 13 judges.

Table 12

Items Identified as Being Susceptible to Test-Wiseness

Identified by Judges:

Items 4, 6, 7, 13*, 15*, 23*, 24*, 26*, 28, 29, 31, 45*, 48*

% agreement (7 or more judges):

54, 69, 85, 85, 92, 69, 54, 69, 77, 62, 69, 85, 69

* Confirmed by empirical evidence from examinees' responses as being test-wise susceptible.

Additional items identified using empirical evidence from examinees' responses:

1, 3, 16, 17, 18, 19, 20, 21, 22, 25, 27, 28, 30, 31, 32, 37, 38, 39, 41, 42, 44, 47, 49

Table 13

Distribution of Test-Wise Items According to Test-Wise Cue Used, Items Showing Chance, and Bad Items

Test-wiseness element[a]			
ID1	Absurd or nonfunctioning distractors/options	19	(63.3%)
ID2	Similar options	0	(0.00%)
ID3	Opposite or different options	4	(13.3%)
ID5	Utilise relevant content information	1	(3.3%)
IIB4	Stem-option link	8	(26.7%)
GUE	Guessing (2, 5, 10, 36, 43, 50)	6	
Bad items (35, 40)		2	

[a]Two items (15 and 27) contained more than one test-wise element.

Table 14

Distribution of Test-Wise Items by Subtest

Subtest	# of items	# of test-wise items	
Mapwork	12	2	(16.6%)
Physical geography	14	9	(64.3%)
Economic geography	19	15	(78.9%)
Population and settlement	5	4	(80.0%)
Total	50	30	(60.0%)

The items identified as being test-wise susceptible by the panel of 13 judges used for this part of the study are listed in the first part of Table 12, together with the percentage agreement among the judges. Although the panel members identified 13 items, the percentage agreement among judges exceeded the 80% standard for claiming agreement for only four. Of these four, empirical evidence obtained from an item analysis of the student responses identified three items: 13, 15, and 45. An additional 23 items were identified from the item analysis results as being test-wise susceptible. Of these, six (items 23, 24, 26, 28, 31, and 48) were identified by at least half of the judges but by less than 80%. Given the problem in judging cognitive thinking level, the empirical evidence was considered to be stronger than judgmental evidence. Consequently, the total number of test-wise-susceptible items was 30, which represents 60% of the total number (50).

Turning to the nature of the test-wise elements identified (see Table 13), the most frequent element was ID1—eliminate options known to be incorrect. Of the 30 test-wise-susceptible items, 19 (63.3%) contain such options. The next most common test-wise-susceptible form was IIB4—stem-option connection—which was found in 8 (26.7%) of the susceptible items, followed by ID3—opposite option—which was found in 4 (13.3%) susceptible items. ID5—utilise relevant content information in other test items and options—appeared once. ID2—similar options—was not observed.

The distribution of items across the test-wiseness elements found in the ZJC geography test is similar to the distributions found by others; for example, Man (1990), Rogers and Bateson (1991), and Rogers and Wilson (1993).

The distribution of test-wise susceptible items across the four topic areas tested by the ZJC geography test is shown in Table 14. With the exception of mapwork, the incidence of test-wise-susceptible items exceeds 60% of the items referenced to each topic.

Returning to the information presented in Table 13, six additional items exhibited an essentially random pattern of response, suggesting that these items were too difficult for the students. An additional two items were badly constructed in that there was more than one correct response.

Evidence of gender differences. Table 15 shows the mean scores for the 414 girls and 495 boys who were in the total sample and the results of the t-tests for independent samples of sex. The mean score for boys, 27.6, was significantly greater ($[t=907] \leq 5.64$; $p < .01$) than the mean for the girls, 25.1. However, the variabilities of the scores for the males and females did not differ significantly ($F[413,494] = 1.84$, nsd; SD) (SD 6.896—girls, 6.356—boys). These findings are consistent with the findings of Elwood (1994) and Stobart et al., (1992), who found that boys do better than girls on a test containing multiple-choice items.

Tables 16 and 17 show a comparison between the performance of boys and that of girls at both the item and subtest levels. Working at the .05 level of significance, at the item level, the boys outscored the girls on 24 (48%) of the 50 items. By contrast, the girls outscored the boys on only 1 (2%) of the 50 items. Although the variability of the scores was very similar for both groups ($F < 1.0$; NS), at the subtest level, the greatest contribution towards the difference in the performance of the boys in comparison to that of the girls came from economic geography, where the mean score for the boys was 1.28 higher than that for the girls. Physical geography

Table 15

t-tests for Independent Samples of Sex

Sex	Mean	SD	SEM	F-value	2-tail prob.	t-value	D.F.	2-tail prob.	t-value	D.F.	2-tail prob.
Girls (414)	25.1053	6.896	0.339	1.18	0.083	-5.64	907	0.000	-5.60	848.78	0.000
Boys (495)	27.5899	6.356	0.286								

SD = Standard deviation
SEM = Standard error of the mean
D.F. = Degrees of freedom

contributed the second largest difference, with a mean score for the boys 0.67 higher than that for the girls, followed by mapwork, with a mean score for the boys 0.31 higher than that for the girls. The smallest contribution towards the mean score differences came from population and settlement, where the boys outscored the girls by 0.22. All the subtest mean score differences were significant at .05 level.

A closer examination of the items where there were significant differences between the performance of boys and that of girls revealed that out of the 24 items where boys outscored girls, 20 items were testing on concepts most likely to be familiar to boys such as agriculture (items 26, 37, 41, 46, 47), industry (items 20, 39), transport (items 11, 19), mining (items 25, 27), settlement design and building (items 9, 48), rock weathering, and runoff (items 13 and 15), and solving realistic problems (items 7, 43).

The only item (36) where girls outscored boys involved solving an abstract problem. This observation agrees with the observation made by Harris and Carlton (1993) that girls seem to do better at solving problems involving abstractions, whereas boys do better at solving real-situation problems.

Table 16

Summary of Differences Between Boys' and Girls' Performance

	Boys		Girls		Difference
Item	p-value	rpbis	p-value	rpbis	p-value (boys-girls)
Mapwork					
1.	47.3	0.37	46.4	0.37	0.9
2.	50.1	0.46	50.0	0.50	0.1
3.	19.8	0.34	19.8	0.23	0.0
4.	46.7	0.45	36.0	0.47	10.7*
5.	16.6	0.07	14.5	0.16	2.1
6.	23.8	0.28	24.9	0.29	-1.1
7.	51.9	0.47	45.2	0.46	6.7*
8.	32.7	0.28	32.1	0.30	0.6
9.	59.6	0.42	51.4	0.40	8.2*
10.	25.1	0.29	26.6	0.33	-1.5
11.	54.9	0.30	47.8	0.31	7.1*
12.	39.8	0.46	42.0	0.44	-2.2
Mean score = 4.68			Mean score = 4.37		0.31
Physical Geography					
13.	75.8	0.43	69.3	0.44	6.5*
14.	53.1	0.33	45.2	0.41	7.9*
15.	83.6	0.42	72.0	0.44	11.6*
17.	20.6	0.30	19.3	0.37	1.3
28.	57.6	0.44	51.0	0.45	6.6*
29.	44.2	0.31	41.3	0.26	2.9
30.	89.7	0.34	85.5	0.37	4.2
31.	46.9	0.36	36.5	0.42	10.4*
32.	72.9	0.45	67.6	0.47	5.3
34.	56.8	0.29	58.9	0.26	-2.1
35.	21.6	0.24	20.8	0.34	0.8
43.	24.6	0.38	16.2	0.34	8.4*
45.	73.3	0.44	74.4	0.41	-1.1
49.	56.8	0.45	53.4	0.51	3.4
Mean score = 7.78			Mean score = 7.11		0.67

(table continues)

Item	Boys p-value	Boys rpbis	Girls p-value	Girls rpbis	Difference p-value (boys-girls)
Economic Geography					
16.	76.0	0.42	71.3	0.38	4.7
18.	84.0	0.47	74.4	0.48	9.6*
19.	82.0	0.38	72.5	0.40	9.7*
20.	76.0	0.34	67.9	0.25	8.1*
21.	46.7	0.28	47.3	0.34	-0.6
22.	69.1	0.24	68.6	0.26	0.5
23.	83.0	0.40	68.1	0.38	14.9*
24.	72.1	0.44	66.4	0.42	5.7
25.	32.7	0.43	20.5	0.36	12.2*
26.	83.4	0.40	73.2	0.44	10.2*
27.	56.6	0.37	43.7	0.41	12.9*
33.	71.7	0.39	66.7	0.42	5.0
36.	23.0	0.00	30.2	0.12	-7.2*
41.	91.9	0.32	85.7	0.38	6.2*
42.	75.6	0.37	67.4	0.28	8.2*
44.	55.4	0.36	51.2	0.21	4.2
46.	55.6	0.37	44.4	0.40	11.2*
47.	52.7	0.43	45.7	0.40	7.0*
50.	20.4	0.25	14.7	0.21	5.7
Mean score = 12.08			Mean score = 10.80		1.28
Population and Settlement					
37.	84.8	0.56	77.8	0.55	7.0*
38.	69.3	0.58	66.9	0.62	2.4
39.	79.6	0.57	70.3	0.67	9.3*
40.	28.1	0.48	32.9	0.40	-4.8
48.	43.4	0.61	34.8	0.57	8.6*
Mean score = 3.05			Mean score = 2.83		0.22

* Denotes items where boys outscored girls (24 items, 48%), significant at .05 level.

\- Denotes items where girls outscored boys (1 item, 2%), significant at .05 level.

Table 17

Gender Differences at the Subtest Level

	Boys = 495		Girls = 414			
	$\bar{X}$	SD	$\bar{X}$	SD	F	t
Mapwork	4.68	1.99	4.37	2.02	0.97_{NS}	2.33_{S}
Physical geography	7.78	2.31	7.11	2.49	0.86_{NS}	4.21_{S}
Economic geography	12.08	2.89	10.80	2.99	0.93_{NS}	6.56_{S}
Population and settlement	3.05	1.22	2.83	1.29	0.89_{NS}	2.65_{S}

NS = Not significant
S = Significant at 0.5 level

Chapter 5

Discussion and Conclusions

This final chapter provides a review of the findings and their interpretation in the light of the theories and issues in research considered in earlier chapters. The first section gives the interpretation of the findings. Implications for practice in Zimbabwe are presented in the second section. Section three looks at the limitations of the study, followed by recommendations for further research in section four.

Interpretation of Findings

Reliability and item/test difficulty. It was implied that in general, with all other things being equal, test reliability increased with the increasing length of a test. In the present study, this observation was found to hold for three of the four subtests that made up the ZJC geography examination. The one exception was the mapwork subtest. Although this subtest contained 12 items, its internal consistency, .39, was lower than the internal consistency for the population and settlement subtest, .46, which contained five items. The main factor accounting for this difference would appear to be the level of difficulty of each subtest. Mapwork was the most difficult subtest, with an average p-value of .38. Population and settlement was easier, with an average subtest p-value of .59. As recently pointed out by Feldt (1993), the internal consistency of multiple-choice tests will be enhanced if the majority of the items have p-values which are concentrated within the vicinity of optimum item difficulty. The main reasons why items with p-values around .50 enhance total test reliability are (a) they maximise observed score variance, and (b) they have the highest levels of discrimination between the upper- and lower-ability groups, which in turn results in higher test reliability. By contrast, the most difficult items (p-values around .05) and the easiest items (p-values around .95) do not differentiate between

ability groups; hence they cannot contribute significantly to the internal consistency of a test.

Feldt (1993) also cautioned against the vigorous pursuit by test developers of items with difficulty values concentrated around .50, because there is no significant advantage to be gained through such an approach. Items with p-values spread from .25 to .75 and a mean close to .50 give similar internal consistency indices to those whose p-values are around .50. Feldt (1993) and Hopkins et al. (1990) also advised that we should not, as test developers, use distractors that represent ultra-fine distinctions in an effort to transform questions on simple ideas into more difficult items. Conversely, it is also not necessary in the interest of test reliability to employ farfetched options to change a question on a normally difficult concept into an easier item.

Some measurement specialists (Ebel & Frisbie, 1991; Lehmann & Mehrens, 1991; Sax, 1989) have observed that although there might be a small reliability advantage for a test with items whose p-values are concentrated around the optimum difficulty level, there may be a genuine psychological advantage for the inclusion of a few easy items at the beginning of a test and a few difficult ones towards the end, with the rest being of average difficulty. Examinees of low and average ability derive greater confidence when they find out that they can answer all of the first five or so items despite having to struggle with the few difficult ones towards the end of the test.

It was also suggested that different methods of calculating reliability indices yield different results. While for a dichotomously scored test, Cronbach's alpha, Hoyt's ANOVA, and K-R20 yield the same value, Hoyt's estimate was .06 higher than Cronbach's composite for alpha (.79 and .73, respectively) because Hoyt's ANOVA is computed at the item level, whereas Cronbach's alpha for composite is computed at the subtest level. Such a finding was consistent with observations made

by Feldt and Brennan (1989): "Stratification on content can result in significantly greater impact" (p. 118). In general, if the stratification is based on item difficulty, there would be little difference between estimates yielded using Cronbach's alpha and Cronbach's alpha for a composite. In practice, however, educational tests are usually stratified by content. By far the more typical test is assembled from items drawn from four or more major content categories as is the case with the 50 items used in the geography test.

Item discrimination and distractor effectiveness. When used in conjunction with p-values, item discrimination indices provide a potent strategy for detecting flawed items, particularly in relation to ambiguity, miskeying, the presence of test-wiseness, and guessing.

An examination of Appendix D on the major psychometric properties of the test items reveals that the trend for examinees to concentrate their responses around a few distractors can be discerned. This trend results from the inclusion in the test of many items which are susceptible to examinees using test-wise strategies together with their partial knowledge to eliminate one or more distractors (test-wiseness is treated separately after this section). The fact that as many as 30 items are susceptible in this test is related to the problem of constructing consistently "good" fourth and fifth distractors. Payne (1992) has observed that there is nothing inherently sacred about the practice of providing four or five distractors for every multiple-choice item used in a test. Similar observations were made by Tversky (1964) and Grier (1971), both of whom were able to show mathematically that the use of three alternatives will maximise the discrimination power and information yielded by item analyses. Such a conclusion makes intuitive sense because we all know how difficult it is to create plausible fourth and fifth distractors. Where these are found, they frequently turn out to be essentially "space fillers."

According to Gronlund (1993), Millman and Greene (1989), and Payne (1992), the main purpose of item analysis is to identify the weaknesses of the items making up a test, so that wherever possible, such items can be either improved or replaced by others as long as the new items come from the same domain. An ideal item, at least from the statistical point of view, is one that all examinees with no relevant knowledge (low-ability group) miss and all examinees with the relevant knowledge (high-ability group) answer correctly. In addition, it would be desirable that the responses of the low-ability group be evenly distributed among the distractors; among the high-ability group, the keyed option should be the most popular. As well, the item discrimination index (point-biserial) should be at least .20, whereas the discrimination index for the foils should be negative. In some testing programmes such as that in the province of Alberta, Canada, a foil is said to be plausible if at least 5% of the examinees select it and the point-biserial is negative.

The most potent strategy for improving item quality is the use of an approach that combines both the conventional item analysis and item response theory. Item analysis is carried out on the fundamental assumption that examinees who exhibit mastery of the subject (high-ability group) in the construct being measured are presumed to be more likely to answer any item about that subject or construct than examinees who exhibit low mastery (low-ability group). Conversely, items that all examinees answer correctly or miss do not discriminate and consequently yield no information about differences between individuals. Such items are not very useful for achievement testing where Norm Referenced Tests (NRTs) are the order of the day because ranking according to individual test scores is the ultimate purpose of most educational tests. However, as pointed out above, some easy items should be included to cater for examinees of low ability.

Content representativeness. One of the fundamental purposes of this study was to assess whether the geography specimen paper for the Zimbabwe Junior

Certificate effectively sampled the curricular specifications contained in the Ministry of Education and Culture (1987b) Zimbabwe Junior Certificate geography syllabus. The specimen paper used in this research was produced by the National Geography Panel, who also were largely responsible for writing the syllabus. It should also be noted that there is only one geography syllabus for the Zimbabwe Junior Certificate level, and every school teaches from this one document according to the stipulations of the Government of Zimbabwe Education Act (1987). It is also pertinent to observe that in the geography syllabus document, it is only the content that is precisely stated. The cognitive skills that are to be tested can be gleaned only from a reading of the whole document and the specimen paper.

In achievement tests similar to the geography test used in this study, the precise specification of content and its validation are perhaps the first stages on the road towards gathering evidence for valid test-score interpretation. A content-validation study seeks to establish a consensus of informed opinions about the degree of congruence between particular test items and specific descriptions of the content area that is intended to be assessed by those items (Messick, 1989; Osterlind, 1989). The panel of judges used to validate the content of the 50 items represented an informed group of opinions, and the results obtained can therefore be regarded as representing a balanced reflection of the content domain. With the exception of 7 out of the 50 items, there was at least 80% agreement across the four topic classifications. The geography test used for the research exercise can thus be viewed as a sound sample of the knowledge acquisition expected from students at the ZJC level.

With respect to classification of the cognitive demands of each item, there was little agreement among the judges. Others (Gierl, 1993; Wolf, Kelson, & Silver, 1990) found similar results.

Multiple-choice tests in particular have been criticised for narrowing the curriculum by focussing on discrete skills. With regard to this criticism, it should be

remembered that any assessment instrument whose test scores are used in important decision-making situations such as selection and certification will automatically become the focus of classroom instruction. Contrary to narrowing the curriculum, multiple-choice items allow for a wider coverage of the content area because of the usually large number of items contained in a single test. In the case of the geography test in question, the lowest number of items per subtest was 5, and the highest was 19. On the other hand, extended test tasks such as essay questions, whilst appearing to call for detailed treatment of the subject matter, are not free from the problem of spotting for the popular examination topics. A cursory glance at a few past examination papers, say, in history, will reveal frequent repetition of particular topics, often with the same wording.

As far as fostering discrete skills is concerned, it is pertinent to observe that any achievement test item is easier constructed around manageable chunks rather than large chunks of a whole body of knowledge. An effective test developer has to have a sense of the "parts" of a discipline that are to be covered in order to ensure that a relatively balanced coverage of the content area is attained (Hambleton & Murphy, 1992). Indeed, it does more good for examinees to know that they are having problems with understanding atmospheric circulation than that they are having problems with understanding the whole of weather or physical geography.

In achievement testing, no single assessment instrument can claim to test all abilities, however well intentioned it may be. And as Lyman (1993) advised, it would not be morally justifiable to use a single test as the sole basis for making a decision. Rather, the tendency is to use at least two or more different instruments to assess the abilities of examinees. If any decision has to be made on the basis of a single test, then that decision ought to be a tentative one, bearing in mind that whilst "tests may reflect or predict ability, they do not cause ability" (p. 40).

The presence of test-wiseness. Judges identified 13 items as being test-wise susceptible. Empirical evidence gleaned from students' responses confirmed only 7 of the 13 items. Empirical evidence from students' performance further identified 23 additional items as being test-wise susceptible, resulting in 30 (60%) of the total test items being classified as being test-wise susceptible.

The finding that most items fell into the absurd functioning distractors (ID1) category confirmed similar findings by Rogers and Bateson (1991) that students do not guess randomly. Rather, the tendency is to eliminate options known to be incorrect and to focus on the remaining few options. In contrast to a student who lacks the relevant knowledge and thus must guess from among all the options, a test-wise student with partial or relevant knowledge will tend to weigh each option until a match is made or guess from among the remaining options. "When a match is made, the cycle is terminated and a test-wise (as opposed to a pure knowledge) response is recorded" (p. 162).

The majority of the examinees were English-as-a-second-language speakers. This situation probably explains why test-wise strategies of stem-option link (IIB4), opposites (ID3), and similar options (ID2) were not common amongst the examinees. Besides, teachers did not coach their students on test-wiseness because the concept is relatively new to the testing situation in the country. Man (1980) and Rogers and Bateson (1991) found that there was a moderate to strong correlation between verbal activity (English language in this case) and test-wiseness. Roznowski and Bassett (1992) observed that test-wiseness responded to coaching. To the extent that test-wiseness refers to an individual's cognitive skills in exploiting the formats and characteristics of the test and/or the test-taking situation to receive a higher score, individuals can be trained or coached to improve their test scores (Roznowski & Bassett, 1992). They also found that extremely difficult items (p-values around .05) and the easiest items (p-values around .80 and .90) responded only slightly or not at

all to coaching. The largest improvements to coaching seem to come from the moderately difficult items (p-values around .50). To the extent that coaching for various score-improvement techniques exists, the validity of inferences drawn from educational tests will obviously be debatable. Rogers and Bateson (1994) have shown that before students can profitably apply test-wise skills, they must first possess some knowledge about the content of the item stem and options. If this knowledge that an examinee possesses is considered relevant to the item content, the interpretation of the total test score as a valid indicator of achievement may be justified. If, however, this knowledge is considered not relevant, the interpretation of the total test score will be confounded by construct-irrelevant easiness (Messick, 1989).

Rogers and Bateson (1991, 1994) also emphasised the role played by experience in multiple-choice testing where experienced test takers tend to exploit flawed items to enhance their total scores, as opposed to inexperienced test takers, who tend to guess randomly. The case of the geography test, where the students had not taken any such examination before, could be viewed as leading to considerable guessing. Besides, multiple-choice testing at this level was only introduced towards the end of the 1980s, and examinees' experience with this kind of examination was still limited. Considerable guessing was also prevalent amongst the more difficult items such as items on mapwork—items 36, 48, and 50—all of which called not only for knowledge but also for elements of application in order to arrive at the correct answer. Such a finding would most probably reflect the emphasis in teaching which still centres around factual recall of knowledge at the relative expense of the higher order cognitive skills.

Gender differences in performance. It was proposed that gender differences in the performance of boys and girls may exist. A review of the relevant literature revealed that (a) differences did exist; (b) in certain subjects (notably mathematics, physics, and languages), initially, the difference favoured boys; but (c) such

differences were decreasing and in some cases (GCSE examinations in the United Kingdom) were being reversed in favour of girls.

The subject of gender differences in performance is a complex one involving ethnicity, class, and equity in addition to gender (Stobart et al., 1992). From the point of view of examination results, the ease with which test scores have been analysed by gender rather than by ethnicity and class has often resulted in differences in performance being attributed solely to gender to the exclusion of other factors. The fact that there was significant difference between the performance of girls and that of boys in the test amounting to 2.5 points (in favour of boys) does not necessarily mean that girls are inherently poor in geography. A number of other factors could be responsible for the difference, as the following will illustrate:

1. The assessment instrument used may favour one sex at the expense of the other. Stobart et al. (1992), Elwood (1994), and Murphy (1994) have observed that a change in the type of assessment instrument used affects performance differently in accordance with out-of-school experiences of the two sexes. They have shown that multiple-choice tests tend to give the advantage to boys, especially if the test encourages considerable guessing, because girls are less likely to take risks than boys are. Similarly, the introduction of a large component of coursework tends to work in favour of girls, who are more likely than boys to work consistently at their assignments.

2. An important change in the whole curricular assessment can result in increased motivation within one sex. An example is cited by Gipps (1994) from Higher School Certificate Physics from South Australia, where girls' performance has improved since 1988 when a female chief examiner was appointed:

> This examiner has deliberately worked within a particular model of Physics (which takes a 'whole view' of the subject); simplified the language of the questions; included contexts only that are integral to particular physics problems; offered a range of different ways of

> answering questions which does not privilege one form of answer over another; provided a list of key instruction words and how students would go about answering questions which include these words. (p. 10)

The fact that boys outperformed girls on the specimen paper considered in this study may be attributed to the fact that 22 of the 24 members of the National Panel that constructed the paper were males. Secondly, the majority of items on economic geography may have favoured boys because the items examined topics which were likely less familiar to girls.

3. The question of equal access to education may have adversely affected girls in that, until after political independence (1980), very few girls entered secondary school (preference was given to boys where a choice between a boy and a girl had to be made). In addition, teenage pregnancy often takes its toll on the 12-13-year-old girl (the rate of expulsions runs at 10% every year at this age group). The net result of the situation cited above means that by comparison there are likely to be more boys of superior quality than there are girls at this level.

Test-score use and interpretation. Throughout this study emphasis has been placed on the need to obtain reliable test scores whose use and interpretation would be meaningful, appropriate, and useful in terms of the specific inferences made from them. The ultimate objective always is to arrive at valid judgments which will result in appropriate decisions and actions with respect to the use of the test scores. In this regard the reporting of students' attainment is central to the purpose of any public examination (Cook, 1991). Unfortunately, it is an area that has not received as much attention as the creation of quality tests and the development and application of valid and reliable marking/scoring procedures.

According to Cole and Moss (1989), any response to a test is a sample of behaviour. Consider the following example of a topographical map-interpretation exercise used in the specimen geography test:

1. What is the height of the land at the dip tank in grid square 0276?

A. 1480 m.
B. 1500 m.
*C. 1520 m.
D. 1540 m.
E. 1560 m.
*Keyed correct answer.

Any examinee's answer to this question is a sample of behaviour. A test score is a summary of several such answers (in this case the test had a total of 50 items). The basic question is, What do these behaviour samples tell us?

If all the questions for simple map interpretation are like this one and a student answers a few correctly, achieving a low score of, say, 15 out of 50, then the obvious conclusion that might be drawn is that the examinee is weak in simple map reading and requires practice in basic map-reading skills. Clearly, this is not the only possibility. It might be that the student knows that contour lines on the map show height but cannot recognise that this is the skill called for in this exercise. Even in a seemingly straightforward map-interpretation exercise such as the one above, the question lends itself to alternative interpretations. The low score might reflect difficulty with the contour interval, with understanding the map key, with map reading as a whole, with motivation, with the teaching method, or with a general fear of anything to do with a practical exercise such as mapwork.

With regard to test-score interpretation, Cronbach (1971, 1988) and Messick (1989) have argued that all test-score interpretations should be examined within an integrated approach to validity. Taken alone, neither construct- nor content- nor criterion-related evidence of validity is sufficient to rule out other plausible rival explanations. In educational measurement the tendency is to test particular persons in a particular setting for particular purposes such as remediation, selection and placement, promotion, certification, evaluating the effectiveness of teaching, focussing the curriculum, and accountability.

Any public examination system must have, as its objective, the provision of information about students based on their examination performance which is consistent with a number of fundamental principles, the most important of which are listed below:

1. recognising the inevitable fallibility of examination test scores or grades and accounting for such measurement error in reporting performances,

2. providing information that is both familiar and adequate to consumers,

3. pitching the information (including statistical information) to the level of understanding of the consumers, and

4. recognising the possibility that current public examinations and their associated credentials may not continue to satisfy the needs of consumers in the context of changing entries.

The common approach of using raw gradings derived from raw scores awarded by examiners (A-F or 1-9 based on marks on a 1-100 scale) often results in a great deal of misclassification or faulty grading, especially where there are fewer points on the scale. The worst situation of misclassification occurs where there are only two points of "Pass/Fail." In fact, it is pertinent to observe here that the reliability of raw grades is less than that on the original scores and, rather than reducing measurement error, actually increases it (Cook, 1991).

Currently, the results of the Zimbabwe Junior Certificate examination are not used for selection or placement. Rather, they are viewed as simply a useful indicator of a student's performance after the first two years of secondary education. If, however, the results are to be used for channelling students into particular streams at the General Certificate of Education (end of fourth-year secondary examination), as has been recently mooted in the country, then serious consideration must be given to the validity of the test scores for this use at this level. To the extent that 60% of the geography specimen paper items were identified as being test-wise susceptible, a

potential validity problem exists when one attempts to interpret the meaning of the test score. Geography may not be the only subject so affected.

One simple and improved method of reporting test scores at the Zimbabwe Junior Certificate level would be to standardise the raw scores before appending grades to them and then to award grades according to standardised-score bands.

Implications for Practice in Zimbabwe

The research study developed from a desire to provide a sound and useful psychometric model upon which to analyse the appropriateness of educational tests in Zimbabwe; and so far from attempting to prove or disprove any theory, the main thrust in the study was to indicate psychometric possibilities for Zimbabwe in the light of existing research experience elsewhere in the world. From the research study six important possibilities were identified. These are summarised below and are recommended for adoption by testing organisations in Zimbabwe.

1. To enhance reliability in a norm-referenced situation like that represented by the JC Geograpy Test, items should be written such that the majority have difficulty values (classical test score model) between .30 and .70, with a small number of easier items so as to encourage lower performing students.

2. In order to improve the extent to which items in a particular test adequately sample a defined domain of knowledge, (a) the specification tables used for constructing educational tests should be validated from time to time using a panel of expert judges as was the case in this study, and (b) test developers should carry out content-validation investigations which focus on the purpose for which the test is to be used, the particular context of testing, and the importance accorded each particular topic.

3. Despite the fact that the judges did not agree at an acceptable level, a similar finding was observed in other jurisdictions that, in order to ensure that more than

pure recall is assessed, a thinking dimension should be included in a table of specifications, and that the item writers have to write items at these levels.

4. In an effort to assess whether test scores from the performance of examinees who took a particular test are a valid indicator of examinee performance, test developers and users should be able to rule out any possible influence of construct-irrelevant easiness arising out of test-wise-susceptible items that are present in the test.

5. To improve the quality of items, especially distractor effectiveness, it is essential that test developers carry out a rigorous item-analysis exercise using both the conventional item analysis and item response theory approaches.

6. In trying to explain gender differences in examinee performance, we should always take into account the influence of other factors such as teachers', parents', and pupils' expectations; examination entry policies (selective or open); type of assessment instrument used; and emphases within the syllabuses, in addition to gender.

7. When interpreting test scores, users of educational tests should take into account the fallibility of test scores arising from measurement error, especially in situations where important decisions such as selection and certificaiton have to be made.

Limitations of the Research

As in all research, this study had its own limitations. The summary conclusions drawn above should therefore be considered in the light of the limitations outlined in this section.

A major limitation of the research study is related to the generalisability of the findings. Because it was not possible to indicate the proportion of the total school types represented by the sample, the findings may not be generalised to the entire

country in a wholesale fashion. Caution is required when applying the findings to the whole country because there may be sample variability.

Another limitation of the study arises from the fact that the model is built on test scores from only one subject and one type of assessment instrument. Much as the framework would be considered illuminative of educational tests in the country, a fuller picture might have been presented by looking at more than one subject area. Distinct disciplines such as science, mathematics, and languages may present different profiles, especially in the area of gender differences.

The absence of corroborative evidence from examinees with regard to their test-taking strategies is considered as one major limitation to the validation of the cognitive skills contained in the table of specifications. Examinees provide a most useful insight into the interpretation of the skills that educational tests purport to test.

As a pioneering study, this research suffered greatly from the absence of any relevant research experiences in the country. It is a well-known axiom that any new research can benefit only from previous studies. Unfortunately, such an opportunity was not afforded this study.

Recommendations for Further Study

The major implication to be drawn from this study is the need for more research in Zimbabwe. All is not perfect on the testing scene in the country. There is a real need for test developers to construct their tests on the basis of firm and sound psychometric principles involving the issues of reliability and validity. This need is made even greater by the realisation that test scores may be used by different consumers for a variety of important decisions.

Arising out of the limitations of the research study outlined above, certain directions for further study are possible. In the first instance there is a need to widen the research base by including more subject areas. The inclusion of science,

mathematics, and languages in further research should enrich the testing experience in the country. It would be interesting, for instance, to find out whether gender differences in performance, which have been identified elsewhere in the world and confirmed for the Zimbabwe Junior Certificate geography, are also present in other subject areas, especially mathematics and English. If the differences are present, how big are they, and is there any indication that the differences are decreasing?

One area requiring further research involves the investigation of specific psychometric properties such as the construct validation of cognitive skills. If the goal of our teaching objectives is to inculcate certain thinking behaviours, it is only fair that protocols are carried out with relevant individual students to determine the actual thinking skills they use. Similarly, if selection is carried out on the basis of test scores from one level (usually lower grade or form), there is a need to establish the predictive validity of these scores as indicators of future performance at higher standards of education.

References

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). (1974). *Standards for educational and psychological tests.* Washington, DC: American Psychological Association.

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Ary, D., Jacobs, L. C., & Razavieh, A. (1990). *Introduction to research in education* (4th ed.). Fort Worth, TX: Holt, Rinehart, and Winston.

Beck, M. D. (1974). Achievement test reliability as a function of pupil-response procedures. *Journal of Educational Measurement, 11*(2), 109-114.

Benson, J. (1988). The psychometric and cognitive aspects of test-wiseness: A review of the literature. In M. H. Kean (Ed.), *Test-wiseness* (pp. 1-25). Bloomington, IN: Phi Delta Kappan.

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (Eds.). (1956). *A taxonomy of educational objectives: Handbook 1: Cognitive domain.* New York: Longman, Green.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3,* 296-322.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Cole, N. S. (1984). Testing and the "crisis" in education: Presidential address presented at the NCME meeting in New Orleans, LA, April 1984. *Educational measurement: Issues and practice, 3*(3), 4-8.

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 201-219). New York: American Council on Education and Macmillan.

Cook, J. S. (1991). Recording public examination performance: Usefulness, accuracy, and ownership. In A. J. M. Luitjen (Ed.), *Issues in public examinations: A selection of the proceedings of the 1990 IAEA Conference (1991)* (pp. 39-44). Utrecht, Austria: Uitgeverij Lemma B.V.

Crehan, K. D., Koehler, R. A., & Slakter, M. J. (1974). Longitudinal studies of test-wiseness. *Journal of Educational Measurement, 11*(2), 209-212.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart, and Winston.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L. J. (1980). Validity on parole: How can we go straight? In W. B. Schrader (Ed.), *Measuring achievement: Progress over a decade (Proceedings of the 1979 ETS Invitational Conference)* (pp. 99-108). San Francisco: Jossey-Bass.

Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Weiner, & H. I. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper and Row.

Cronbach, L. J., Schonemann, P., & McKie, D. (1965). Alpha coefficients for stratified-parallel tests. *Educational and Psychological Measurement, 25,* 291-312.

Ebel, R. L. (1961). Must all tests be valid? *American Psychologist, 16,* 640-647.

Ebel, R. L., & Frisbie, D. A. (1986). *Essentials of educational measurement* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Elwood, J. (1994, April). *Undermining gender stereotypes: Examination performance in the UK at 16.* Unpublished paper presented at the 1994 AERA/NCME Annual Conference, New Orleans, LA.

Feingold, A. (1988). Cognitive gender differences are disappearing. *American Psychologist, 43,* 95-103.

Feldt, L. S. (1993). The relationship between the distribution of item difficulties and test reliability. *Applied Measurement in Education, 6*(1), 37-48.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 105-146). New York: American Council on Education and Macmillan.

Foxman, D., Ruddock, G., & McCallum, I. (1990). *APU mathematics monitoring 1984-88 (Phase 2).* London: School Examinations and Assessment Council.

Gierl, M. J. (1993). *Evaluating Bloom's cognitive levels in the table of specifications for a Grade 6 mathematics achievement test.* Unpublished master's thesis, Department of Educational Psychology, University of Alberta, Edmonton, AB.

Gipps, C. (1994, April). *What do we mean by equity in relation to assessment?* Unpublished paper presented at the 1994 AERA/NCME Annual Conference, New Orleans, LA.

Goldstein, H. (1994). Recontextualizing mental measurement. *Educational Measurement: Issues and Practice, 13*(1), 16-19, 43.

Government of Zimbabwe. (1987). 1987 Education Act. Harare, Zimbabwe: Government Printer.

Grier, J. B. (1975). The number of alternatives for optimum test reliability. *Journal of Educational Measurement, 12*(2), 109-113.

Gronlund, N. E. (1991). *How to write and use instructional objectives* (4th ed.). New York: Macmillan.

Gronlund, N. E. (1993). *How to make achievement tests and assessments* (5th ed.). Needham Heights, MA: Allyn and Bacon.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10,* 255-282.

Hambleton, R. K., & Murphy, E. (1992). A psychometric perspective on authentic measurement. *Applied Measurement in Education, 5*(1), 1-16.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Harris, A. M., & Carlton, S. T. (1993). Patterns of gender differences on mathematics items on the Scholastic Aptitude Test. *Applied Measurement in Education, 6*(2), 137-133.

Harris, M. M., & Pickle, J. G. (1992). Creating an equitable environment: Gender equity in Lincoln, Nebraska. *The Educational Forum, 57* (Fall), 12-17.

Hopkins, K. D., Stanley, J. C., & Hopkins, B. R. (1990). *Educational and psychological measurement and evaluation* (7th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Howarth, R. B. (1991). *Assessment of national educational goals and objectives: A case study of the localisation of ordinary level examinations in Zimbabwe.* Unpublished paper presented to the 17th International Conference of the International Association of Educational Assessment (IAEA), Nairobi, Kenya.

Hoyt, C. J. (1941). Test reliability estimated by analysis of variance. *Psychometrika, 6,* 153-160.

Hyde, J. S. (1981). How large are cognitive gender differences? A meta-analysis using w^2 and d. *American Psychologist, 36*(8), 892-901.

Jacklin, C. N. (1989). Female and male: Issues of gender. *American Psychologist, 44*(2), 127-133.

Kelley, T. L. (1942). The reliability coefficient. *Psychometrika, 7,* 75-83.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3,* 635-694 (Monograph Supplement).

Lord, F. M. (1952). A theory of test scores. *Psychometric Monograph No. 7.*

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Lumsden, C. A. (1976). Test theory. *Annual Review of Psychology, 27,* 251-280.

Lyman, H. B. (1991). *Test scores and what they mean* (5th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Maccoby, E. E., & Jacklin, N. C. (1974). *The psychology of sex differences.* Stanford, CA: Stanford University Press.

Maguire, T. O., Hattie, J., & Haig, B. (1993, June). *Construct validity and achievement assessment.* Unpublished paper presented at the Cognition and Assessment Conference of Canada at Queen's University, Kingston, ON.

Man, D. W. (1990). *Cross cultural study of test-wiseness.* Unpublished master's thesis, Department of Psychology and Special Education, University of British Columbia, Vancouver, BC.

Masango, R. B., & Nembaware, L. (1991). *The implementation of national goals and objectives in Zimbabwe and replacement of University of Cambridge Local Examinations Syndicate's International 'O' Level Syllabuses by local syllabuses.* Unpublished paper presented at the 17th International Association of Educational Assessment (IAEA), Nairobi, Kenya.

Mehrens, W. A. (1984). National tests and local curriculum: Match or mismatch? *Educational Measurement: Issues and Practice, 3*(3), 9-15.

Mehrens, W. A., & Lehmann, I. J. (1987). *Using standardized tests in education* (4th ed.). New York: Longman.

Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology* (4th ed.). New York: Holt, Rinehart, and Winston.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Weiner & H. Braun (Eds.), *Test validity* (pp. 33-45). Hillsdale, NJ: Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). New York: American Council on Education and Macmillan.

Millman, J., Bishop, C. H., & Ebel, R. L. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement, 25*(3), 707-726.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 335-366). New York: American Council on Education and Macmillan.

Ministry of Education and Culture. (1987a). *The curriculum development unit plan (1986-1990).* Harare, Zimbabwe: Author.

Ministry of Education and Culture. (1987b). *Zimbabwe junior certificate geography syllabus: Curriculum development unit.* Harare, Zimbabwe: Author.

Moss, P. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational research, 62*(3), 229-258.

Murphy, P. (1989). Assessment and gender. *National Union of Teachers Education Review, 3,* 37-41.

Murphy, P. (1994, April). *Sources of inequity: Understanding students' responses to assessment.* Unpublished paper presented at the 1994 AERA/NCME Annual Conference, New Orleans, LA.

Nhandara, E. S. (1993). *Programme evaluation design: The localisation of ordinary level examinations in Zimbabwe.* Unpublished project design submitted in partial fulfillment of the requirements of Educational Administration 515, University of Alberta, Edmonton, AB.

Nhandara, E. S., & Chagonda, T. T. T. (1990, October). *Processing of examination results from marking to the time they are released.* Unpublished paper presented at the Conference of Heads of Examinations Institutions in East and Central Africa, Lusaka, Zambia.

Osterlind, S. J. (1989). *Constructing test items.* Norwell, MA: Kluwer Academic Publishers.

Payne, D. A. (1992). *Measuring and evaluating educational outcomes.* New York: Macmillan.

Principles of fair student assessment practices for education in Canada. (1993). Edmonton, AB: Joint Advisory Committee, Centre for Research in Applied Measurement and Evaluation, University of Alberta.

Rogers, W. T., & Bateson, D. J. (1991). The influence of test-wiseness on performance of high school seniors on school-leaving examinations. *Applied Measurement in Education, 4*(2), 159-183.

Rogers, W. T., & Bateson, D. J. (1994). Verification of a model of test-taking behavior of high school seniors. *Alberta Journal of Educational Research, 40*(2), 195-211.

Rogers, W. T., & Wilson, C. (1993). *The influence of test-wiseness upon performance of high school students on Alberta Education's Diploma Report.* Unpublished report submitted to Alberta Education Student Evaluation Branch, Edmonton, AB.

Roznowski, M., & Bassett, J. (1992). Training test-wiseness and flawed item types. *Applied Measurement in Education, 5*(1), 35-48.

Rulon, P. J. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review, 9,* 99-103.

Sammons, P. (1994, April). *Gender and ethnic differences in attainment and progress: A longitudinal analysis of student achievement over nine years.* Unpublished paper presented to the 1994 AERA/NCME Annual Conference, New Orleans, LA.

Sarnaki, R. E. (1979). An examination of test-wiseness in the cognitive domain. *Review of Educational Research, 49*(2), 252-279.

Satterly, D. (1989). *Assessment in schools* (2nd ed.). Oxford, UK: Basil Blackwell.

Sax, G. (1980). *Principles of educational and psychological measurement and evaluation* (2nd ed.). Belmont, CA: Wadsworth.

Sax, G. (1989). *Principles of educational and psychological measurement and evaluation.* Belmont, CA: Wadsworth.

Shepard, L. A. (1991). Will national tests improve student learning? *Phi Delta Kappan* (November), 232-238.

Smith, J. K. (1982). Converging on correct answers: A peculiarity of multiple-choice items. *Journal of Educational Measurement, 19*(3), 211-219.

Smith, M. L., & Glass, G. V. (1987). *Research evaluation in education and the social sciences.* Englewood Cliffs, NJ: Prentice-Hall.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15,* 72-101.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3,* 271-295.

Stobart, G., Elwood, J., & Quinlan, M. (1992). Gender bias in examinations: How equal are the opportunities? *British Educational Research Journal, 18*(3), 263-276.

Stufflebeam, D. L., Foley, W. J., Gephart, W. J., Guba, E. G., Hammond, R. L., Morrison, H. O., and Provus, M. M. (1971). *Educational evaluation and decision making.* Itasca, IL: Peacock.

Tversky, A. (1964). On the optimal number of alternatives at a choice point. *Journal of Mathematical Psychology, 1,* 386-391.

Wolf, A., Kelson, M., & Silver, R. (1990). *Learning in context: Patterns of skills transfer and training implications.* London: Training Agency.

Appendix A

Glossary of Educational Measurement Terms

Appendix A

Glossary of Educational Measurement Terms

The section contains some 31 educational assessment terms whose meaning is considered to be so special as to deserve some explication or definition.

1. *Assessment:* Assessment is the systematic evaluative appraisal of an individual's ability and performance in a particular environment or context. In educational measurement, assessment has been used to refer to all the strategies and techniques that might be used to collect information from students about their progress toward attaining the knowledge, skills, attitudes, or behaviours prescribed in an instruction/learning programme. In practice, common assessment strategies and instruments include observations, dialogue, individual tasks, group projects, paper/pencil/pen tests, portfolios, informal/formal surveys, peer evaluation, and creative writing.
2. *Evaluation* suggests that a judgment must be made regarding what constitutes worth or value. Typically, the term is associated with how effective or ineffective, how adequate or inadequate, how good or bad, how valuable or invaluable, and how appropriate or inappropriate a given action, process, or product is in terms of the perceptions of the individual who makes use of information provided by an evaluator (assessment instruments). Indeed, as Stufflebeam et al. (1971) correctly observed, the purpose of evaluation is to improve, not to prove.
3. *Measurement* is the process of assigning numbers according to some specified rule such that the numbers reflect differences in the amount of the variable possessed by different individuals or units. The numbers can give the extent of the difference on an ordinal, quasi-interval, interval, and ratio scale.
4. *Ability* as applied to measurement refers to mental power, native or acquired, that enables one to do things well. Ability can also be applied generally to

assessment when it refers to competence or aptitude for a particular kind of action.

5. *Achievement testing:* An achievement test is a test which is designed to measure how much a child has acquired as the result of specific teaching. The acquisitions are usually of information, knowledge, or skills which schools have made a deliberate intention to teach.
6. *Psychometric* is a psychological term used to refer to the measurement of traits, abilities, and processes.
7. *Test score* refers to the alphanumeric grade or percentage assigned to represent the ability (performance) of an individual in a particular achievement test.
8. *Unidimensionality,* which is common to all educational tests, is an assumption that an item or test is measuring a single latent ability or trait which in turn affects examinee performance. In practice, however, it is impossible to achieve total unidimensionality, because examinee performance is obviously influenced by more than one latent ability. The tendency therefore in the identification of unidimensionality is to recognise dominant components or factors which are then referred to as the ability measured by the test. From the psychometric point of view, unidimensionality is reflected in the total test reliability coefficients and the correlations between subtests and total test.
9. *Test item* refers to a single item, task, or exercise in a test. Examples of test items are a single multiple-choice item, a single practical task, a single essay question, or a single mathematical calculation. A test item in an examination of mental attributes is a unit of measurement with a stimulus and a prescriptive form for answering, and it is intended to yield a response from an examinee from which performance in some psychological construct (such as ability, predisposition, or trait) may be inferred.

10. *Sub test* refers to a component or part of a total test. In multiple-choice testing, a single item constitutes a subtest, as would combinations of items on the same topic such as mapwork or economic themes.
11. *Total test* refers to the aggregate of all the items that must be answered in order to constitute a full test.
12. *Item analysis* refers to a set of methods for examining an item to determine its difficulty, discrimination, and validity. Item-analysis procedures also allow test constructors to discover items that are either ambiguous, miskeyed, or susceptible to the notion of test-wiseness.
13. *Item difficulty:* The difficulty level of an item (p) is the proportion (percentage) of students responding correctly to it.
14. *Item discrimination* indices (D) measure the extent to which items differentiate between those persons with the highest and lowest scores on the total test (normally, the upper 27% and lower 27%, leaving the middle 46%). The highest-scoring persons within a cohort of students taking a particular test are the upper-criterion group; the lowest-scoring persons on the same test are the lower-criterion group. Items are negatively discriminating if a larger proportion of students in the lower group do better on an item than do those in the upper group; in such cases the D index is then negative.
15. *Biserial and point-biserial correlations* which are given in many item-analysis computer outputs are essentially a measure of how effectively an item discriminates or differentiates between examinees who are relatively high on the ability level and those who are relatively low (in other words, those who have the knowledge and those who do not). A more detailed practical application of the biserial and point-biserial correlation is given in the chapter on methodology.

16. *Item ambiguity:* An item is said to be ambiguous to the extent that students in the upper group select an incorrect option with about the same frequency as they select the correct one.
17. *Miskeying:* The possibility of miskeying is shown when a large number of students in the upper group respond to what the test constructor believes is a distractor (that is, the wrong answer).
18. *Test-wiseness* is a cognitive ability or set of skills that a test taker can use to improve a test score no matter what the content area of a test is. If a test taker possesses test-wiseness and if the test contains susceptible items, then a combination of these two factors can result in an improved score; in contrast, a student low in test-wiseness will tend to be penalised every time he or she takes a test that includes test-wise components. Thus, a potential validity problem exists when one attempts to interpret the meaning of the test score.
19. *Dichotomous scoring system* refers to the scoring system which divides the scoring into two—right or wrong.
20. *Polychotomous scoring system* describes a system of scoring that includes several scoring systems at least over and above right or wrong.
21. *Reliability* refers to the extent to which a test is consistent or dependable in measuring whatever it is that it measures. In one sense, therefore, reliability is the degree to which test scores are free from errors of measurement. According to classical test theory, the test score (observed score) is made up of two parts: the true score (τ) + error of measurement (E).
22. *Standard error of measurement* is a statistic applied to a single obtained score to estimate the amount by which that score differs from the hypothetical "true score." The larger the standard error, the greater the amount of error contained in the score. In statistical manipulations the standard error of measurement is

the standard deviation of the test times the square root of one minus the reliability coefficient.

23. *Reliability coefficient* is the coefficient of correlation between two forms of a test, between scores on two administrations of the same test taken a week or so apart, or between halves of the same test after Spearman-Brown correction. Other important types of reliability coefficients are provided by the Kuder-Richardson formula 20 and Cronbach's alpha, which are both based on the total items in the test. Hoyt's estimate of reliability is based on the analysis of item variance and yields the same coefficient as KR-20 or Cronbach's alpha. Cronbach's alpha for a composite is computed on the basis of the subtest reliabilities.

24. *Standard deviation* is a linear measure of the "scatter" or "dispersion" of scores in a distribution, obtained by finding the square root of the size of the average squared deviation (variance) from the mean. In a normal distribution, approximately two thirds (68.3%) of the scores occur within the range from one standard deviation below the mean to one standard deviation above the mean.

25. *Validity* refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. Test validation is the process of accumulating evidence to support inferences made from test scores. To date, such evidence is divided into three broad fields of construct-, content-, and criterion-related evidence.

26. *Construct-related evidence* focusses primarily on the test score as a measure of some psychological or theoretical construction of interest. Reasoning ability, spatial visualisation, reading comprehension, sociability, introversion, endurance, and leadership are all examples of constructs which could be measured using specific conceptual frameworks that distinguish one from the other.

27. *Content-related evidence* generally demonstrates the extent to which the sample of items, tasks, or questions on a test are representative of some defined universe or domain of content.
28. *Criterion-related evidence* demonstrates that test scores are systematically related to one or more outcome criteria. In the case of test scores, useful criterion-related evidence could be gathered by referring to the performance of the upper and lower criterion groups.
29. *Validity generalisation* refers to the degree to which scientific conclusions can be drawn from validity studies, on the one hand; and the use of the results of validity evidence obtained from prior studies to support the use of a test in a new situation, on the other hand. The latter use raises fundamental questions about the degree to which validities are transportable to a specific new situation.
30. *Differential prediction* is a broad concept hat includes the possibility that different prediction equations may be obtained for different demographic groups, for groups that differ in their prior experiences, or for groups that receive different treatments or are involved in different instructional programmes. The term *treatment* is intended to include not only the various forms of intervention, but also the manner in which tests are administered, such as by computer.
31. *Specification table* refers to a two-way grid (or table) summarising the objectives of teaching and assessment. Entries in the grid specify the importance of the various topics, the relative importance to be allotted to them in the construction of the achievement test, and the various skills that will be tested.

Appendix B

Zimbabwe Junior Certificate Geography Specimen Paper

STUDY THE MAP EXTRACT OF HEADLANDS ON THE SCALE OF 1:50 000.
QUESTIONS 1-12 REFER TO THIS MAP.
ALL QUESTIONS MUST BE ANSWERED ON THE GRID PROVIDED

1. What is the height of the land at the dip tank in grid square 0276?

   A 1480 m.

   B 1500 m.

   C 1520 m.

   D 1540 m.

   E 1560 m.

2. Which one of the following squares contains the steepest slopes?

   A 0380.

   B 0280.

   C 0180.

   D 0080.

   E 9980.

3. In the area to the north and east of the main tarred road which one of the following occupies the largest area?

   A Cultivation.

   B Dense bush .

   C Forest plantation.

   D Medium bush.

   E Sparse bush .

4. What is the name given to the point marked 1524•54 in grid square 0377?

   A A bench mark .

   B A trigonometrical station.

   C A spot height.

   D A boundary beacon.

   E A contour line.

5. An area of map is shown in figure 1.1 below :

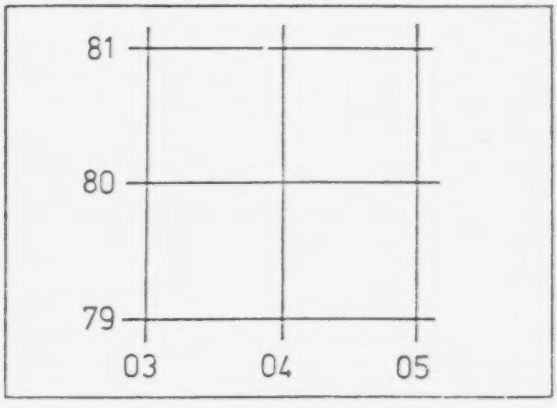

Figure 1.1 Part of the map grid

In this area the river Mwarazi is flowing

A slowly to the S.W.

B quickly to the S.W.

C slowly to the N.E.

D quickly to the N.E.

E quickly to the N.W.

6. The main lines of communication shown run

A uphill from N.W. to S.E.

B in a valley between two upland areas.

C on flat ground from S.E. to N.W.

D uphill from S.E. to N.W.

E uphill from N.E. to S.W.

7. What is the distance from the dip tank in grid square 0079 to the rifle range building in grid square 9876?

A 3.0 km.

B 3.5 km.

C 7.0 km.

D 4.5 km.

E 4.0 km.

8. What is the bearing of the quarry in grid square 0477 from the building in York Estate in grid square 0079?

A 024°.

B 100°.

C 295°.

D 065°.

E 115°.

9. The settlement pattern of huts in grid square 9878 is

   A clustered in the same place.
   B dispersed among the fields.
   C linear along a footpath.
   D circular around a source of water.
   E radial along several footpaths.

10. How does the drainage to the south of the railway line differ from that to the north?

   A To the north there is more man-made drainage.
   B To the south the rivers are flowing faster.
   C To the north they are cutting deeper valleys.
   D To the south the rivers flow north-west.
   E To the south there are more rapids.

11. Using map evidence what is the main activity of the people who live outside Headlands?

   A Transport and administration.
   B Ranching and growing crops.
   C Mining and industry.
   D Tourism and entertainment.
   E Fishing and forestry.

12. What is the grid reference of spot height 1540?

   A 016769.
   B 021806.
   C 008766.
   D 026769.
   E 016767.

13. The type of weathering caused by the expansion and contraction of rocks due to heating and cooling is called

   A erosion.
   B exfoliation.
   C oxidation.
   D solution.
   E decomposition.

14. Study figure 1.2 below and then answer the question that follows:

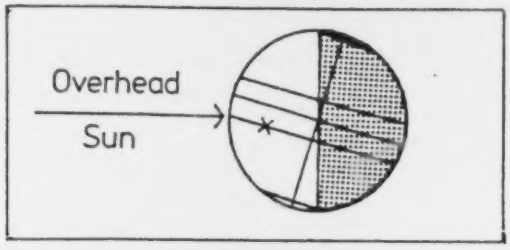

Figure 1.2. Revolution of the earth

During this time of the year, the season at point X is

A winter.
B between winter and spring.
C autumn.
D summer.
E spring.

15. The term RUNOFF refers to water

A which flows on the earth's surface and forms rivers.
B which is stored in a very large dam.
C which goes into the ground to form water.
D vapour which condenses in order to form clouds.
E which is used for the irrigation of crops.

16. Which one of the following is NOT a tertiary activity?

A Teaching.
B Banking.
C Insurance.
D Mining.
E Telecommunications.

17. Examine figure 1.3 below and then answer the question that follows:

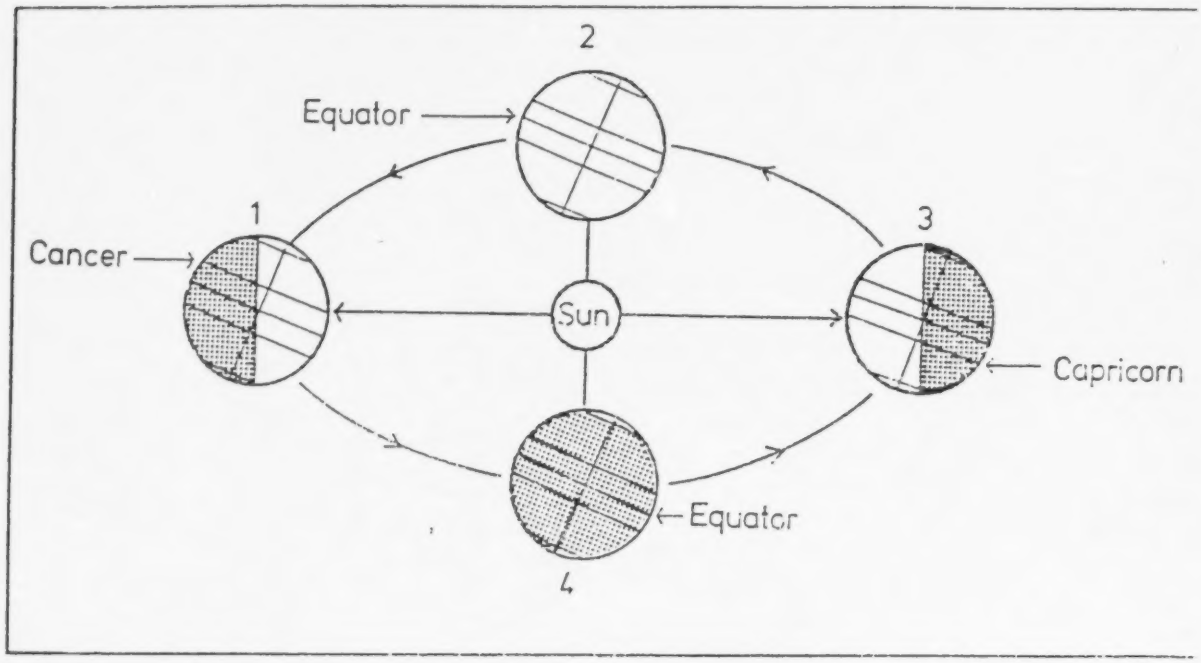

Figure 1.3 Day and Night.

Which of the following correctly shows the area in darkness on June 21?

A position 1.
B position 2.
C position 3.
D position 4.
E none of these positions.

18. Which one of the following is NOT a water conservation method?

A Sharing bath water with a friend.
B Recycling used water.
C Collecting rainwater falling on rooftops.
D Washing clothes once a week.
E Cleaning cars using a hosepipe.

19. In Zimbabwe, which one of the following means of transport is the cheapest and capable of carrying bulky loads?

A Water.
B Air.
C Road.
D Rail.
E Pipeline.

20. Which one of the following industries and products is NOT correctly paired?

	INDUSTRY	PRODUCT
A	Food.	Mealie-meal.
B	Metal products.	Window frames.
C	Timber.	Tables.
D	Paper and printing.	Books.
E	Textiles.	Boots.

21. Which one of the following groups of products is worth transporting by air between Zimbabwe and Britain?

A Maize, cotton bales.
B Copper, nickel.
C Medicine, emeralds.
D Motor cars, motor cycles.
E Bricks, cement.

22. Mhangura mine is important for the mining of

A copper.
B lead.
C gold.
D asbestos.
E iron ore.

23. Imports are

A goods sold by a country to another.
B ports handling a variety of goods.
C goods produced and sold within a country.
D goods bought by one country from another.
E always perishable goods moved quickly.

24. Heavy unmoving traffic during peak periods is known as traffic

A census.
B flowline.
C lights.
D accidents.
E congestion.

25. Which one of the following is NOT a raw material used in the manufacture of steel?

A Iron ore.

B Copper.

C Limestone.

D Coal.

E Chrome.

26. Study Figure 1.4. below:

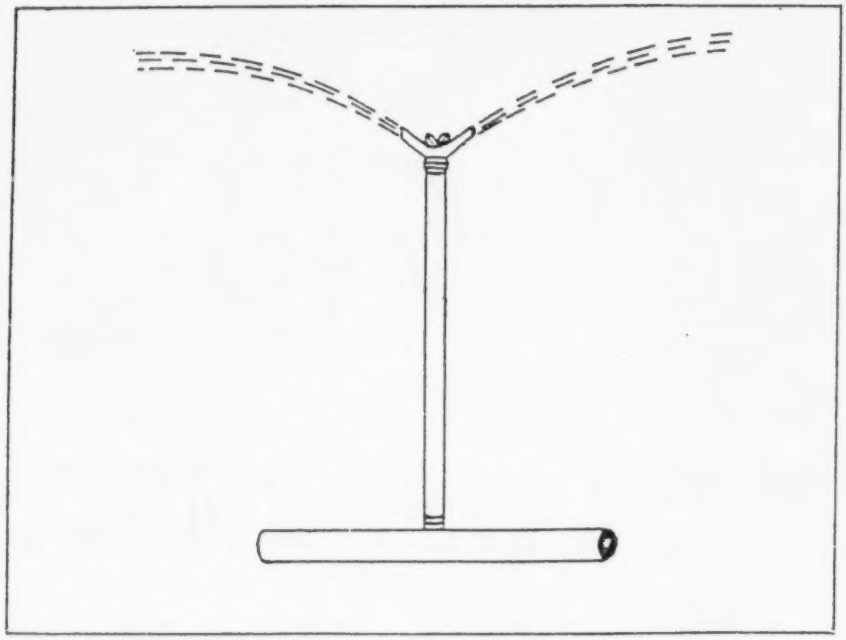

Figure 1.4 An irrigation method

The type of irrigation shown by Figure 1.4. is known as the

A flooding system.

B siphoning system.

C channel system.

D sprinkler system.

E shadoof system.

27. Shaft mining is used when the mineral ores are

A close to the surface and in a large deposit.

B close to the surface and in a small deposit.

C deep down and in a large deposit.

D found on the surface of a river bed.

E in an exposed deposit at the side of a hill.

28. Study the climatic graphs (figure 1.5.) of Harare and Cape Town shown below.

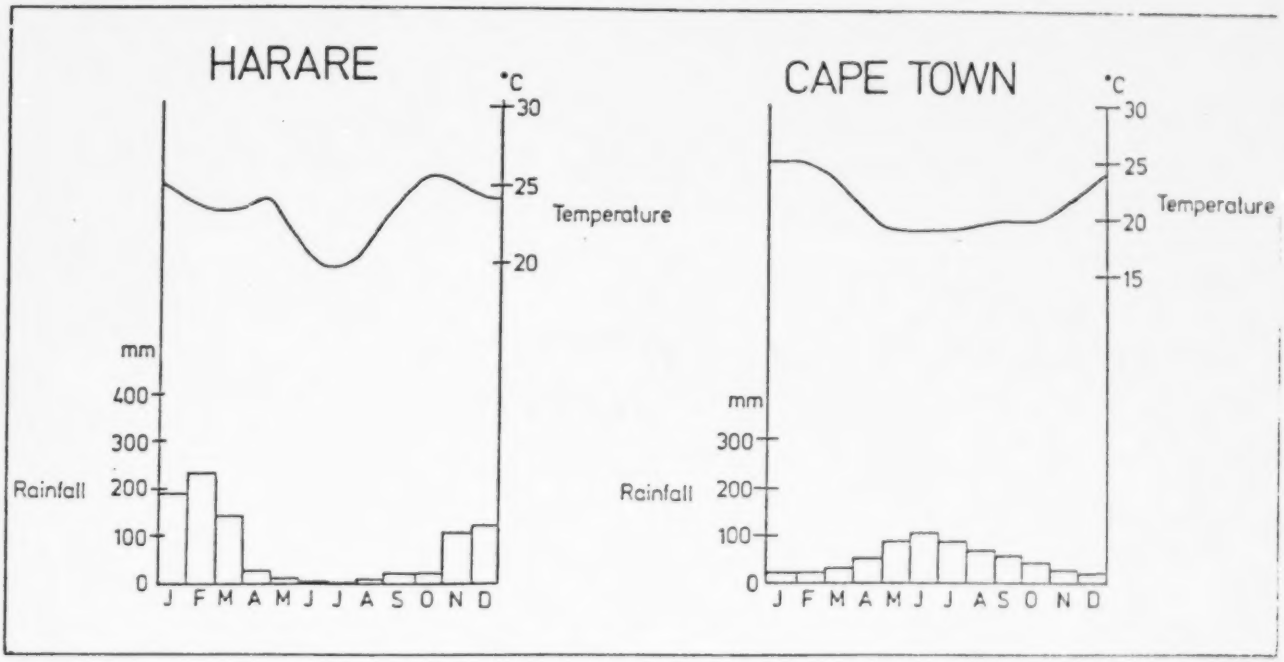

Figure 1.5. Climatic graphs for Harare and Cape Town.

The graphs indicate that the lowest temperatures are found in the season of

A lowest rainfall in Harare but not in Cape Town.

B highest rainfall in both Harare and Cape Town.

C highest rainfall in Harare but not in Cape Town.

D lowest rainfall in Harare.

E lowest rainfall in Cape Town.

29. Look again at figure 1.5. Which one of the following types of natural vegetation would grow BEST around Cape Town?

A Broad-leaved equatorial rainforest.

B Tall grass savanna parkland.

C Mediterranean type evergreen forest.

D Semi-desert thorn bushes.

E Tropical monsoon forest.

30. Study the map (Figure 1.6) which shows annual rainfall in Zimbabwe.

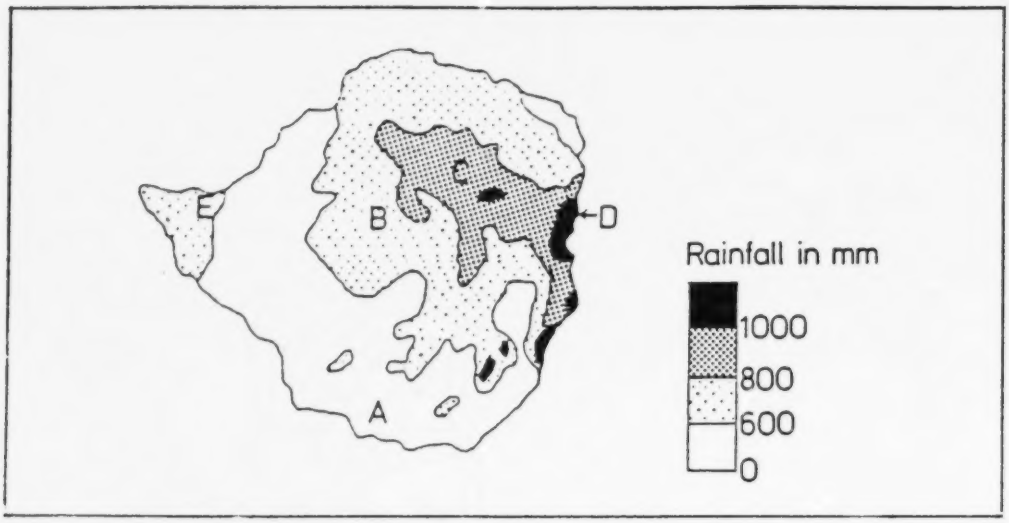

Figure 1.6. Annual rainfall in Zimbabwe.

Which place, A,B,C,D or E on the map receives the lowest total annual rainfall?

31. From point C on figure 1.6, which of the following statements BEST describes the map?

A Rainfall increases to the north.

B Rainfall increases to the west.

C Rainfall decreases to the south.

D Rainfall is the same in all directions.

E Rainfall increases to the south west.

32. A wet and dry bulb thermometer is used to measure

A ground temperature.

B relative humidity.

C maximum daily temperature.

D minimum daily temperature.

E wind speed.

33. Which one of the following energy sources is renewable?

A Solar energy.

B Coal.

C Nuclear energy.

D Petroleum.

E Natural gas.

34. The baobab tree found in some parts of Zimbabwe is suited to the savanna climate because it has

A narrow leaves.

B thin coarse bark.

C a large swollen trunk.

D flexible branches.

E long thorns.

35. The diagram below (figure 1.7.) shows recordings made at a weather station.

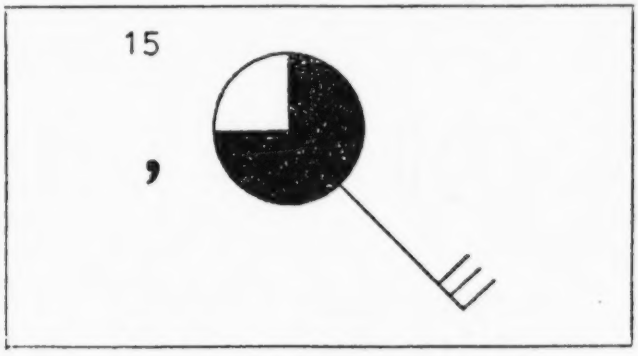

Figure 1.7 Weather station recording

What was the weather like on this day?

A Slight breeze, north-west wind and temperature of $15^{o}C$.

B Strong breeze, south-east wind and drizzle.

C Strong breeze, north-west wind and $\frac{2}{8}$ cloud cover.

D Slight breeze, south-east wind and drizzle.

E Slight breeze, temperature of $15^{o}C$ and $\frac{6}{8}$ cloud cover.

36. The diagram below (figure 1.8) shows 5 possible sites for a new iron and steel factory.

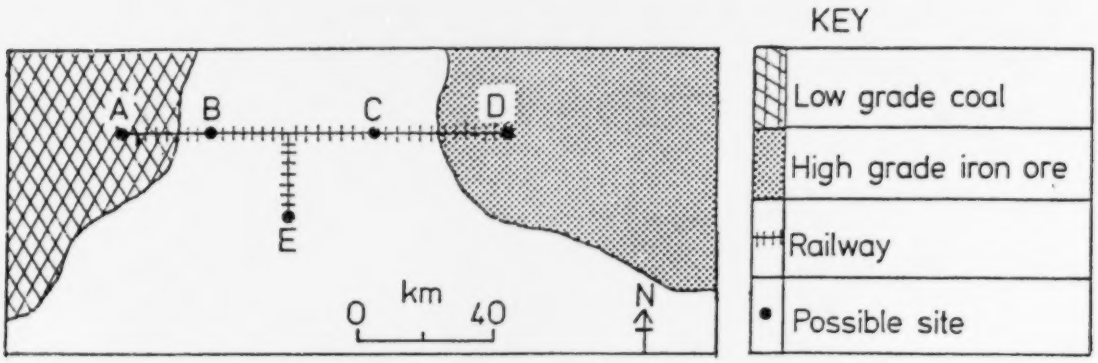

Figure 1.8 Siting an iron and steel factory.

The production of steel at this factory requires half as much iron ore as coal. The costs of moving coal and iron are the same. Which site, A,B,C,D or E will have the LOWEST total transport costs?

37. Which one of the following has had the BIGGEST effect in reducing infant death rates in Zimbabwe?

A Changing farming methods.

B Improvement of housing.

C Expansion of secondary schools.

D Primary Health Care programmes.

E Bringing industries to the rural areas.

38. Which one of the following would you ALWAYS expect to see in the Central Business District (CBD) of a large city?

A Office buildings.

B Residential housing.

C Heavy industry.

D Municipal parks.

E Market gardens.

39. A clothing manufacturer wishes to open a factory in a country. She is given the population information shown in figure 1.9. below.

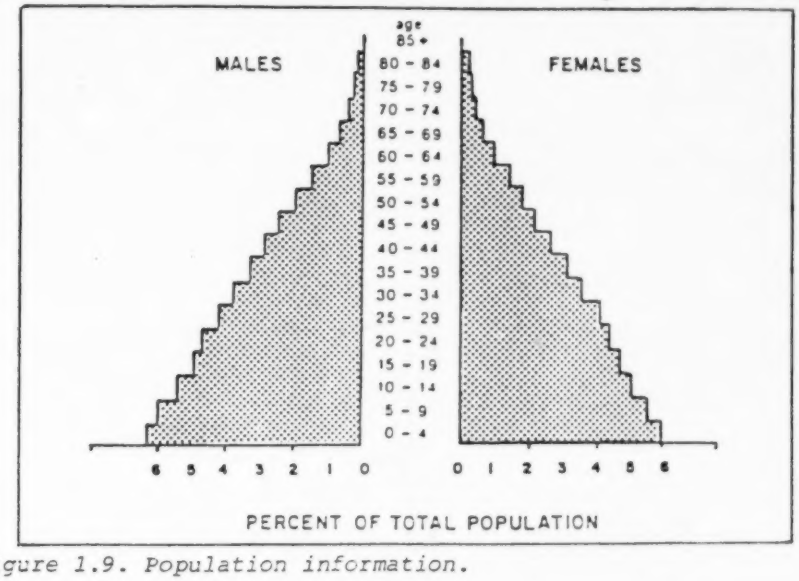

Figure 1.9. Population information.

What type of clothes are in most demand?

A Clothes for people over 80.

B Clothes for people between 65-79.

C Clothes for people between 40-64.

D Clothes for people between 25-39.

E Clothes for people under 25.

40. Look again at figure 1.9. What IMPORTANT information is NOT given to the manufacturer?

A Age structure of population.

B Increase or decrease of population.

C Sex structure of population.

D Geographical distribution of population.

E Dependancy ratio of population.

41. Market gardening is the intensive

A rearing of poultry for eggs.

B growing of vegetables, fruits or flowers.

C growing of tobacco, cotton or maize.

D rearing of dairy cows for milk.

E growing of rice in paddy fields.

42. Study the table below (figure 1.10) which shows the percentage use of energy from different sources for the years 1875 and 1985.

Energy Source	Year	
	1875	1985
Wood	60%	13%
Coal	38%	27%
Oil	2%	40%
Natural Gas	0%	15%
Nuclear energy.	0%	5%

Figure 1.10. Changing use of world energy (%)

Which of the fuels has declined the MOST in importance since 1875?

A Coal.
B Nuclear energy.
C Oil.
D Natural gas.
E Wood.

43. In which of the groups listed below are ALL the rocks sedimentary?

A Shale, limestone, coal.
B Schist, basalt, shale.
C Quartzite, limestone, coal.
D Granite, sandstone, coal.
E Limestone, dolerite, shale.

44. Which one of the following crops is BEST suited to low rainfall areas?

A Maize.
B Rice.
C Wheat.
D Tea.
E Sorghum.

✓

45. Study figure 1.11 below which shows the sketch sections of several landforms.

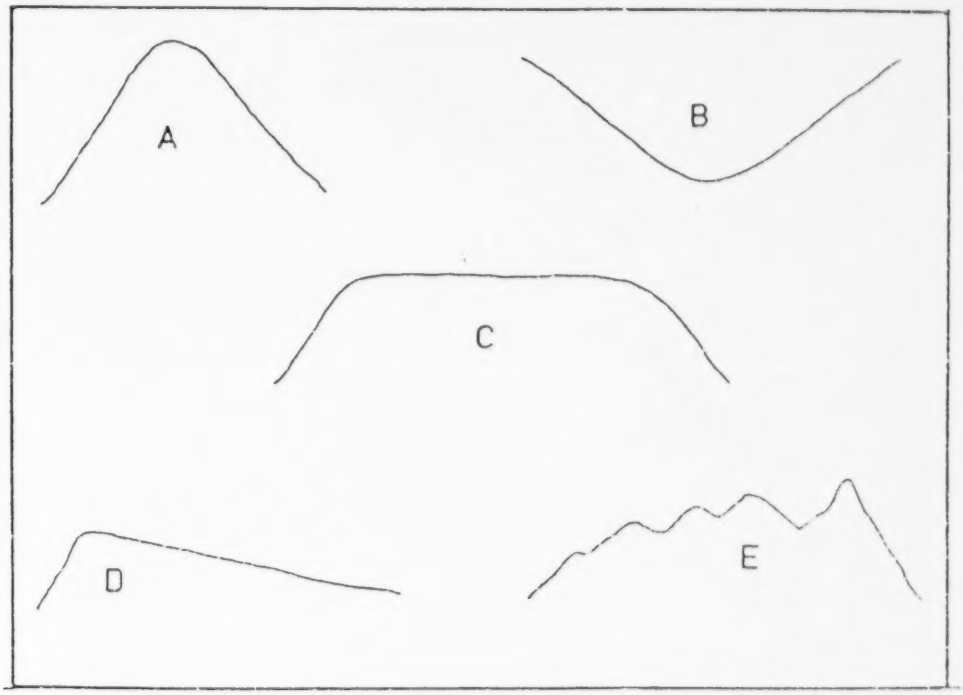

Figure 1.11 Sketch sections of landforms

Which one of the drawings A,B,C,D or E is a sketch section of a conical hill?

46. All of the following statements explain the location of dairy farms near towns and cities EXCEPT

A transport costs are lower.

B milk is delivered whilst it is fresh.

C the best grazing land is found here.

D markets are near these areas.

E veterinary services are close by.

47. From the following, select a good method of conserving soils.

A Growing crops along river banks.

B Planting along contours.

C Cultivating down slopes.

D Growing the same crop yearly.

E Growing crops in vleis.

48. Study the settlement pattern shown below in figure 1.12.

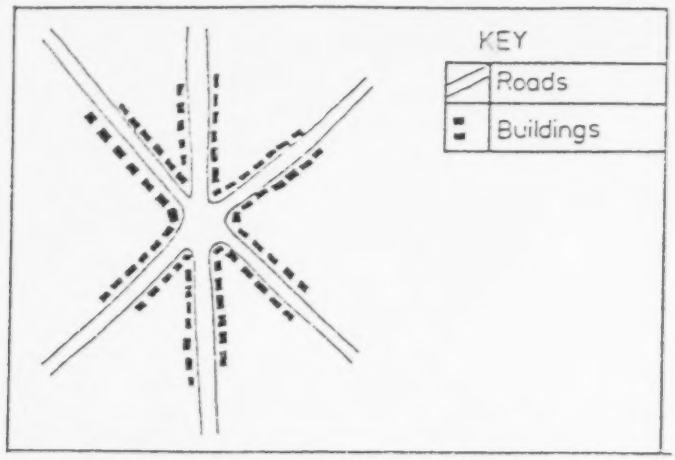

Figure 1.12. A settlement pattern.

The settlement pattern is

A circular.
B linear.
C rectangular.
D haphazard.
E radial.

49 Study figure 1.13. below.

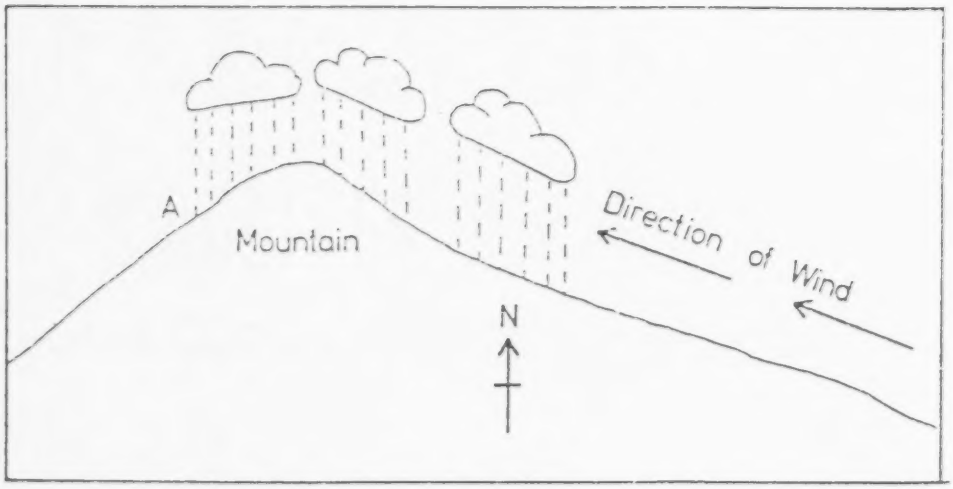

Figure 1.13. Sketch section of a mountain.

On which side of the mountain does A lie?

A Windward.
B Eastern.
C Leeward
D Wind Gap.
E Southern.

50. Study the diagram below (Figure 1.14).

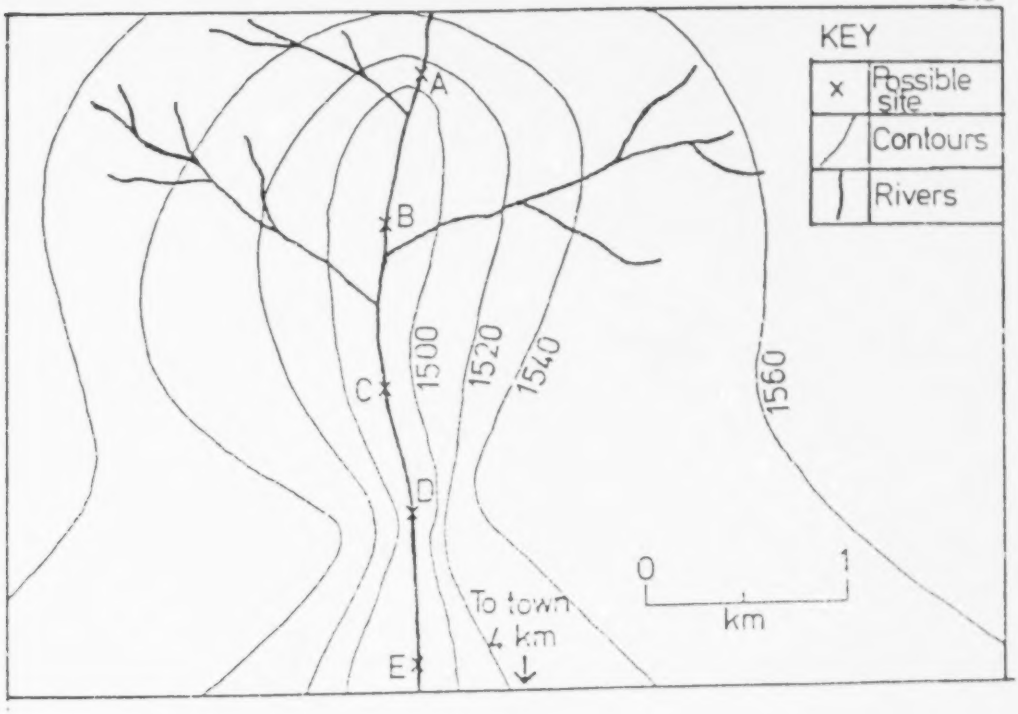

Figure 1.14. Possible sites for a dam wall.

A new dam is needed to increase the supply of drinking water to a nearby town. Which site, A,B,C,D, or E on the map is MOST suitable to build the dam wall.

ANSWER SCHEME FOR MULTIPLE CHOICE

1.	C	26.	D
2.	A	27.	C
3.	E	28.	A
4.	A	29.	C
5.	E	30.	A
6.	D	31.	C
7.	B	32.	B
8.	E	33.	A
9.	B	34.	C
10.	C	35.	B
11.	B	36.	A
12.	A	37.	D
13.	B	38.	A
14.	D	39.	E
15.	A	40.	D
16.	D	41.	B
17.	A	42.	E
18.	E	43.	A
19.	D	44.	E
20.	E	45.	A
21.	C	46.	C
22.	A	47.	B
23.	D	48.	E
24.	E	49.	C
25.	B	50.	D

Appendix C

Instruction Sheet to Judges

Zimbabwe Junior Certificate Geography Specimen Paper

Specifications Table

Appendix C

Instruction Sheet to Judges

Task (a): Classifying items into topic and skill categories:

Please classify each one of the 50 multiple-choice items into the relevant cell provided on the specification grid. Each item should be placed into ONLY one cell and according to your judgment. Note that the classification involves identifying the topic to which the item belongs and recognising the major skill that the item is testing. Some guide notes are also included to help you interpret the cognitive domain being assessed by each item (Bloom et al.'s taxonomy).

Task (b): Identifying items susceptible to the notion of test-wiseness:

More than 10 questions out of the 50 clearly contain elements of test-wiseness. This means that they are phrased in such a way as to (a) give away the answer by providing cues to the correct option, and (b) have one, two, three, or four distracters which are so obviously wrong as to allow candidates to concentrate on four, three, two, or one possible answer. According to Messick (1989), Millman et al. (1965), and Rogers and Bateson (1991), the common flaws contained in multiple-choice items and exploited by test-wise candidates have to do with the four elements listed below:

1. Three deductive-reasoning strategies:
   - ID1 - Eliminate options known to be wrong (non-functioning distracters).
   - ID2 - Choose neither or both of two options which imply the correctness of each other (similarities).
   - ID3 - Choose neither or one of two options, one of which, if correct, would imply the incorrectness of the other (opposites).
   - ID5 - Utilise relevant content information in other test items and options.
2. One-cue-using strategy:
   - IID4 - Recognise and use similarities between the stem and the options.

Please identify the items that appear susceptible to the notion of test-wiseness.

Zimbabwe Junior Certificate Geography Specimen Paper

(Multiple-Choice Paper)

		Question classification by topic			
		Mapwork based on the 1:50000 map extract of Headlands	Physical geography based on weather, climate, water cycle, and land features	Economic geography based on water, energy, farming, mining, manufacturing, and transport	Population and settlement studies
Question classification by knowledge and skill objective	Knowledge with understanding of generalizations, key concepts, and facts in geography				
	Comprehension of information and ideas related to geographical data presented in questions				
	Skills (their recall and application)				
	TOTAL				

Appendix D

Major Psychometric Properties of the Test Items

Appendix D

Major Psychometric Properties of the Test Items

Table 16

Major Psychometric Properties of the Test Items

Item	Option	p-value	rpbis	Test-wise element	Comment (either on TTW or guessing)
Mapwork					
1.	a	5.3	-0.05	ID1 (E)	Using contour lines, the correct answer is given by the contour line passing closest to the dip tank. E lies at the farthest distance from the dip tank.
	b	7.5	-0.07		
	c*	46.9	0.20		
	d	36.1	-0.12		
	e	3.5	-0.05		
2.	a*	50.1	0.41	(GUE)	Applying the concept of a steep slope to contour lines proved difficult for 50% of the students, who tended to guess across options B, C, D, and E.
	b	14.4	-0.18		
	c	9.6	-0.19		
	d	12.5	-0.15		
	e	13.1	-0.12		
3.	a	36.6	-0.05	1D3 (A)	Option A is different from all others; "smarter" students selected E. A, B, C, and D are all plausible distractors for all candidates.
	b	13.9	-0.10		
	c	13.6	-0.21		
	d	15.5	-0.00		
	e*	19.8	0.21		
4.	a*	41.8	0.37		
	b	20.2	-0.05		
	c	18.1	-0.03		
	d	9.4	-0.26		
	e	12.2	-0.23		
5.	a	19.6	-0.03	(GUE)	A very difficult item requiring application, synthesis, and analysis in order to arrive at the correct answer. More examinees settled for distractors A, B, and C than for key E, one of the two items with a negative correlation.
	b	27.9	-0.11		
	c	22.7	-0.04		
	d	13.4	-0.01		
	e*	15.6	-0.03		
6.	a	15.8	-0.00		
	b	20.0	-0.11		
	c	20.6	-0.06		
	d*	24.2	0.28		
	e	18.2	-0.03		
7.	a	14.0	-0.12		
	b*	48.8	0.32		
	c	15.5	-0.12		
	d	14.0	-0.05		
	e	7.7	-0.09		

(table continues)

Item	Option	p-value	rpbis	Test-wise element	Comment (either on TTW or guessing)
8.	a	18.8	0.02		
	b	8.9	-0.15		
	c	19.8	-0.04		
	d	19.5	-0.09		
	e*	32.5	0.29		
9.	a	8.6	-0.12		
	b*	55.9	0.34		
	c	16.0	-0.13		
	d	13.1	-0.19		
	e	6.3	-0.09		
10.	a	18.8	-0.05	(GUE)	A difficult question which led to a lot of guessing. Options B and E, although similar, are not working.
	b	23.8	-0.04		
	c*	25.7	0.15		
	d	12.2	-0.06		
	e	17.7	-0.03		
11.	a	9.7	-0.11		
	b*	51.7	0.24		
	c	9.0	-0.13		
	d	9.8	-0.03		
	e	19.6	-0.10		
12.	a*	40.8	0.35		
	b	9.1	-0.18		
	c	17.1	-0.20		
	d	17.6	-0.12		
	e	13.9	0.02		
Hoyt's estimate of reliability = 0.39 for 12 items Standard error of measurement = 2.01 Mean score = 4.54					
Phys. geog. 13.	a	5.9	-0.22	ID1 (D)	Option D (solution) is the only one which is not directly related to the stem terms *expansion and contraction* and *heating and cooling*, thereby rendering it an absurd option.
	b*	72.8	0.40		
	c	10.6	-0.16		
	d	1.8	-0.16		
	e	8.9	-0.20		
14.	a	21.6	-0.09		
	b	13.2	-0.22		
	c	9.4	-0.07		
	d*	49.5	0.28		
	e	5.7	-0.03		
15.	a*	78.3	0.39	ID3 (B)	Option B is opposite in meaning to key A. *Runoff* literally means to flow.
	b	1.2	-0.09	IIB4 (A)	
	c	7.2	-0.13		
	d	7.4	-0.30		
	e	5.7	-0.16		

(table continues)

Item	Option	p-value	rpbis	Test-wise element	Comment (either on TTW or guessing)
17.	a*	20.0	0.23	ID1 (B, E)	The better students selected A. Option B shows no darkness, and E is the only place where the distractor type is used.
	b	5.3	-0.02		
	c	27.3	0.08		
	d	40.0	-0.23		
	e	6.8	-0.04		
28.	a*	54.6	0.42	ID5 (A, D)	This is a bad item where there are two correct answers, A and D, but one subsumes the other.
	b	6.8	-0.17		
	c	15.3	-0.20		
	d	15.6	-0.09		
	e	7.4	-0.22		
29.	a	12.9	-0.11		
	b	20.1	-0.06		
	c*	42.9	0.16		
	d	16.3	-0.04		
	e	7.7	-0.01		
30.	a*	87.8	0.32	ID1 (B, C, D, E)	From point C on the map, the only sensible rainfall distribution that can be deduced is "a decrease southwards"; the rest are working against the interpretation of the key.
	b	3.1	-0.14		
	c	1.9	-0.17		
	d	4.1	-0.20		
	e	2.6	-0.11		
31.	a	19.9	-0.11	ID1 (D)	From the map, key D is an absurd option.
	b	15.2	-0.12		
	c*	42.1	0.30		
	d	6.4	-0.01		
	e	16.3	0.16		
32.	a	7.9	-0.12	ID1 (D, E)	Wind speed is not measured by a thermometer. Humidity affects people when there is high temperature; this would explain the popularity of option C. A is a plausible distractor to all those students with partial knowledge.
	b*	70.5	0.42		
	c	15.7	-0.31		
	d	3.6	-0.07		
	e	2.0	-0.22		
34.	a	17.3	0.06		
	b	9.2	-0.05		
	c*	57.8	0.13		
	d	7.2	-0.18		
	e	8.3	-0.11		
35.	a	9.7	-0.13	Bad item	D and E are plausible distractors.
	b*	21.2	0.17		
	c	8.1	-0.16		
	d	12.9	0.06		
	e	47.6	-0.02		

(table continues)

Item	Option	p-value	rpbis	Test-wise element	Comment (either on TTW or guessing)
43.	a*	20.8	0.32	(GUE)	A difficult item where examinees showed a lot of "misinformation."
	b	8.0	-0.01		
	c	30.5	-0.09		
	d	20.3	-0.15		
	e	11.1	-0.08		
45.	a*	73.8	0.39	IIB4 (A)	The shape of the sketch section in option A is clearly a cone, whereas other options are depressions, a plateau, or an escarpment.
	b	6.1	-0.19		
	c	6.6	-0.14		
	d	4.1	-0.12		
	e	9.2	-0.22		
49.	a	25.0	-0.16	ID1 (B, D, E)	Options B and E refers to the cardinal points of the compass, but E has nothing to do with the direction of winds shown in Figure 1.13. Option D does not describe a mountainside.
	b	11.4	-0.24		
	c*	55.2	0.40		
	d	5.7	-0.14		
	e	2.4	-0.13		
Hoyt's estimate of reliability = 0.55 for 14 items Standard error of measurement = 1.56 Mean score = 7.47					
Economic geog.					
16.	a	6.2	-0.16	ID1 (A, B, C)	Activities A, B, and C are likely to be better known to students than is E.
	b	1.9	-0.05		
	c	5.6	-0.12		
	d*	73.8	0.36		
	e	12.4	-0.25		
18.	a	4.1	-0.21	ID1 (A, B, C, D)	Easy item which gave away the correct option, E.
	b	5.8	-0.26		
	c	7.2	-0.27		
	d	3.1	-0.12		
	e*	79.6	0.48		
19.	a	6.5	-0.15	ID1 (B, E)	Air, road, and rail transport are currently the only forms of transport applicable to Zimbabwe. B and E are absurd opposites.
	b	2.8	-0.18		
	c	11.0	-0.19		
	d*	77.8	0.34		
	e	1.9	-0.10		
20.	a	12.7	-0.09	ID1 (C, D)	C and D are well known to students.
	b	8.8	-0.12		
	c	4.4	-0.13		
	d	1.7	-0.09		
	e*	72.3	0.24		
21.	a	17.6	-0.04	ID1 (E)	Bricks or cement (option E) are too bulky and heavy to be transported by air.
	b	12.2	-0.10		
	c*	47.0	0.21		
	d	19.5	-0.13		
	e	3.6	-0.04		

(table continues)

Item	Option	p-value	rpbis	Test-wise element	Comment (either on TTW or guessing)
22.	a*	68.9	0.25	IIB4 (A)	*Mhangura* is the Shona word for copper.
	b	2.0	-0.00		
	c	8.0	-0.08		
	d	13.4	-0.17		
	e	7.4	-0.00		
23.	a	17.4	-0.27	ID1 (B)	Option B is the only one that talks about ports; the rest refer to goods being handled in a variety of ways.
	b	0.7	-0.16		
	c	5.1	-0.14		
	d*	76.2	0.36		
	e	0.7	-0.09		
24.	a	4.4	-0.16	IIB4 (E)	Option E, congestion, literally describes unmoving traffic.
	b	19.9	-0.27		
	c	4.0	-0.16		
	d	2.0	-0.15		
	e*	69.5	0.43		
25.	a	2.0	-0.06	ID1 (A)	Wherever it occurs, iron ore is always associated with the manufacture of steel. In fact, most geography books/literature always refer to "iron and steel" as one industry.
	b*	27.2	0.38		
	c	22.6	-0.13		
	d	25.0	-0.19		
	e	23.1	-0.05		
26.	a	3.4	-0.21	IIB4 (D)	The diagram, Figure 1.4, shows water being forced out under pressure in the form of a spray or sprinkle.
	b	5.1	-0.15		
	c	8.7	-0.21		
	d*	78.8	0.40		
	e	3.7	-0.15		
27.	a	20.6	-0.15	IIB4 (C)	A shaft is always associated with considerable depth.
	b	11.7	-0.18	ID1 (D)	
	c*	50.7	0.39		
	d	4.4	-0.16		
	e	12.2	-0.13		
33.	a*	69.4	0.35		
	b	13.2	-0.24		
	c	4.7	-0.11		
	d	5.3	-0.17		
	e	7.2	-0.07		
36.	a*	26.3	-0.06	(GUE)	A very difficult item where students tended to guess, especially between options A, C, D, and E.
	b	12.8	0.15		
	c	12.4	0.06		
	d	17.6	0.01		
	e	30.1	-0.10		

(table continues)

Item	Option	p-value	rpbis	Test-wise element	Comment (either on TTW or guessing)
41.	a	1.9	-0.10	IIB4 (B)	Market gardening is linked to the growing of vegetables, flowers, and fruits for sale.
	b*	89.1	0.31		
	c	5.1	-0.19		
	d	2.8	-0.14		
	e	1.1	-0.15		
42.	a	9.4	-0.13	ID3 (D)	According to the table of information, natural gas shows an increase in use instead of a decline.
	b	7.7	-0.08		
	c	10.1	-0.16		
	d	0.9	-0.11		
	e*	71.8	0.27		
44.	a	8.1	-0.09	ID3 (B, E)	Growing conditions for rice are known to be different from or the opposite of those for sorghum; that is, "wet" versus "dry" conditions.
	b	4.1	-0.10		
	c	17.3	0.00		
	d	16.8	-0.14		
	e*	53.5	0.19		
46.	a	16.6	-0.13		
	b	6.9	-0.11		
	c*	50.5	0.32		
	d	8.0	-0.07		
	e	17.6	-0.18		
47.	a	4.3	-0.08	ID1 (D)	Monoculture (option D) is always associated with soil exhaustion.
	b*	49.5	0.37		
	c	15.0	-0.21		
	d	3.5	-0.10		
	e	27.4	-0.16		
50.	a	9.9	-0.16	(GUE)	A difficult item involving map interpretation. Better students chose D or E; some chose C.
	b	14.0	-0.16		
	c	20.6	0.01		
	d*	17.8	0.22		
	e	37.3	0.11		
Hoyt's estimate of reliability = 0.61 for 19 items Standard error of measurement = 1.83 Mean score for subtest = 11.50					
Population and settlement					
37.	a	4.2	-0.25	IIB4 (D)	Link between health and life.
	b	6.6	-0.20		
	c	1.2	-0.14		
	d*	81.6	0.47		
	e	6.2	-0.27		
38.	a*	68.2	0.39	ID1 (D)	In Zimbabwe municipal parks are closely associated with residential areas rather than with Central Business Districts.
	b	9.9	-0.18		
	c	13.0	-0.21		
	d	1.9	-0.05		
	e	6.8	-0.19		

(table continues)

Item	Option	p-value	rpbis	Test-wise element	Comment (either on TTW or guessing)
39.	a	6.6	-0.28	ID1 (A, B)	Options A and B represent older and fewer people who also tend to be conservative in their clothing requirements—a poor market for a clothing manufacturer.
	b	3.1	-0.22		
	c	5.1	-0.25		
	d	9.6	-0.14		
	e*	75.4	0.47		
40.	a	7.4	-0.16	Bad item	A bad item with two plausible distractors, B and E.
	b	20.7	-0.01		
	c	8.1	-0.14		
	d*	30.3	0.18		
	e	32.8	0.05		
48.	a	14.2	-0.18	ID1 (C)	The shape of the settlement pattern is definitely not rectangular, so many students avoided this option.
	b	24.1	-0.06		
	c	3.9	-0.17		
	d	18.0	-0.19		
	e*	39.5	0.40		

Hoyt's estimate of reliability = 0.46
Standard error of measurement = 0.83 Mean score for subtest = 2.95

* Keyed option.

Appendix E

An Outline of Test-Wiseness Principles

Appendix E

An Outline of Test-Wiseness Principles

I. Elements independent of test constructor or test purpose

A. Time-using strategy

1. Begin a work as rapidly as possible with reasonable assurance of accuracy.
2. Set up a schedule for progress through the test.
3. Omit or guess at items (see I.C. and II.B.) which resist a quick response.
4. Mark omitted items, or items which could use further consideration, to assure easy relocation.
5. Use time remaining after completion of the test to reconsider answers.

B. Error-avoidance strategy

1. Pay careful attention to directions, determining clearly the nature of the task and the intended basis for response.
2. Pay careful attention to the items, determining clearly the nature of the question.
3. Ask examiner for clarification when necessary, if it is permitted.
4. Check all answers.

C. Guessing strategy

1. Always guess if right answers only are scored.
2. Always guess if the correction for guessing is less severe than a "correction for guessing" formula that gives an expected score of zero for random responding.
3. Always guess even if the usual correction or a more severe penalty for guessing is employed, whenever elimination of options provides sufficient chance of profiting.

D. Deductive-reasoning strategy

1. Eliminate options which are known to be incorrect, and choose from among the remaining options.
2. Choose neither or both of two options which imply the correctness of each other.
3. Choose neither or one (but not both) of two statements, one of which, if correct, would imply the incorrectness of the other.
4. Restrict choice to those options which encompass all of two or more given statements known to be correct.
5. Utilize relevant content information in other test items and options.

II. Elements dependent upon the test constructor or purpose

A. Intent-consideration strategy

1. Interpret and answer questions in view of previous idiosyncratic emphases of the test constructor in view of the test purpose.
2. Answer items as the test constructor intended.
3. Adopt the level of sophistication that is expected.
4. Consider the relevance of specific detail.

B. Cue-using strategy

1. Recognize and make use of any consistent idiosyncrasies of the test constructor which distinguish the correct answer from incorrect options:
   a. He makes it longer (shorter) than the incorrect options.
   b. He qualifies it more carefully, or makes it represent a higher degree of generalization.
   c. He includes more false (true) statements.
   d. He places it in certain physical position among an ordered set of options (such as the middle of the sequence).
   e. He places it in a certain logical position among an ordered set of options (such as the middle of the sequence).
   f. He includes (does not include) it among similar statements, or makes (does not make) it one of a pair of diametrically opposite statements.
   g. He composes (does not compose) it of familiar or stereotyped phraseology.
   h. He does not make it grammatically inconsistent with the stem.
2. Consider the relevancy of specific detail when answering a given item.
3. Recognize and make use of specific determiners.
4. Recognize and make use of resemblances between the options and an aspect of the stem.
5. Consider the subject matter and difficulty of neighboring items when interpreting and answering a given item. (Millman et al., 1965, p. 707)